PRÉTRAITEMENT SEGMENTAL ET RECONNAISSANCE DE LA PAROLE EN MILIEU BRUITÉ

Jean-Baptiste PUEL, Régine ANDRÉ-OBRECHT

IRIT - URA 1399 CNRS
Université Paul Sabatier - 118, route de Narbonne
31062 Toulouse cedex - France

RÉSUMÉ

Les systèmes de reconnaissance de parole actuels offrent de bonnes performances pour des applications indépendantes du locuteur sur de petits vocabulaires. Cependant, les performances de tels systèmes dépendent pour une grande part de l'adéquation entre les conditions d'apprentissage et leurs conditions réelles d'utilisation. Nous proposons un ensemble de prétraitements visant à atténuer l'influence des conditions d'enregistrement et de transmission du signal sur les performances des systèmes de reconnaissance : une segmentation automatique du signal, un détecteur bruit/parole et un algorithme de débruitage sont utilisés. Nous présentons les corpus sur lesquels ont porté nos expériences, ainsi que les résultats obtenus.

ABSTRACT

Speech recognition systems offer generally good results for speaker independent applications on small vocabularies. However, such system performances depend for a great part on the similarity between the learning conditions and the real use conditions. We propose some preprocessing methods aiming to lower the mismatch between the learning data and the exploitation data: automatic segmentation, noise/speech detection and noise substraction algorithm are used.
We present the corpus on which we test the algorithms, the results we obtain and we compare the importance of each part of the preprocessing.

1. INTRODUCTION

Dans le cadre de la reconnaissance automatique de parole indépendante du locuteur, les Modèles de Markov Cachés permettent de reconnaître de petits vocabulaires avec de très bonnes performances, y compris dans des conditions d'utilisation difficiles (réseau téléphonique, conditions de bruit automobile, etc...).
Cependant, pour obtenir de bonnes performances en reconnaissance, les conditions d'apprentissage doivent être le plus proche possible des conditions réelles d'utilisation. Dans certains cas, il peut être contraignant, voire impossible d'enregistrer les corpus d'apprentissage dans les conditions où la reconnaissance sera réalisée.
Nous proposons donc un ensemble de prétraitements visant à s'affranchir de cette contrainte : l'objectif étant de conserver des performances correctes lors de la reconnaissance, même si les conditions de bruit sont très différentes de celles rencontrées en phase d'apprentissage.
La première partie de cette étude expose comment le signal est analysé avant la phase de reconnaissance - une segmentation automatique est effectuée ainsi que l'étiquetage des segments en termes de bruit ou parole - et ensuite adapté en fonction du bruit rencontré : une soustraction spectrale du bruit est réalisée.
Nous évaluons dans la seconde partie ce que cette approche apporte à un système classique de reconnaissance à base de Modèles de Markov Cachés. Les expérimentations ont porté sur différents corpus (Esprit ARS et corpus CNET) et permettent d'analyser l'apport respectif de chaque module du prétraitement.

2. PRÉTRAITEMENT

Les module de prétraitement sont utilisés aussi bien lors de l'apprentissage des modèles que lors de la reconnaissance : dans les deux cas, les données subiront un ou plusieurs des traitements suivants.
La segmentation automatique est réalisée systématiquement, et peut être utilisée seule. L'algorithme de détection bruit/parole permet éventuellement de réaliser des systèmes de reconnaissance sans modélisation du bruit. L'algorithme de soustraction spectrale utilise les résultats des deux modules précédents, comme l'illustre la figure 1. Tous ces traitements sont réalisés "en ligne", en temps réel - l'ensemble des paramètres de chacun des algorithmes étant évalué pour chaque nouvelle occurence.

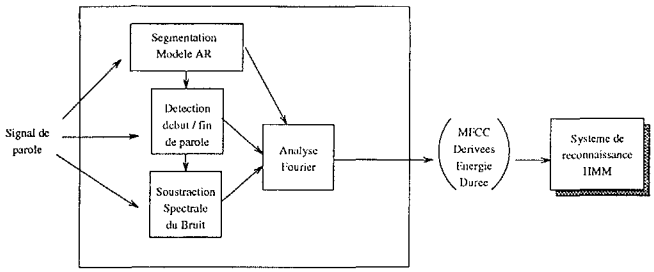

FIG. 1 - : Les modules de prétraitement

2.1. La segmentation automatique

L'algorithme de segmentation utilisé est la méthode de "Divergence Forward-Backward" [André-Obrecht 88], qui permet de localiser les zones quasi-stationnaires de signal. Ce premier module du prétraitement est effectué systématiquement: les observations utilisées par le système de reconnaissance sont limitées à un vecteur par segment. De plus, l'information apportée par la position des ruptures dans le signal est utilisée par les autres modules. Notons enfin que cet algorithme est indépendant des conditions d'enregistrement et du locuteur.
On considère que le signal est représenté par une chaîne d'unités homogènes, chacune d'entre elles représentée par un modèle AR, $\lambda = (a_1, \cdots, a_p, \sigma)$. La méthode consiste à détecter les changements dans les paramètres du modèle. Le test est basé sur le contrôle d'une distance entre les deux modèles λ_0 et λ_1 localisés comme sur la figure 2.

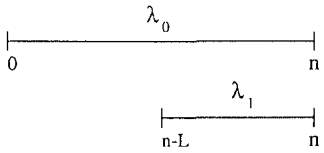

FIG. 2 - : Localisation des deux modèles

Les segments obtenus peuvent être rangés dans trois catégories :

- des segments stationnaires correspondant aux parties stables du signal,
- des segments de transition dans lesquels on trouve une structure formantique dont le comportement reste monotone,
- des segments courts (de l'ordre de 10 ms) qui correspondent à des changements articulatoires rapides, comme l'explosion d'une plosive.

L'expérience a montré que cet algorithme de segmentation détecte toutes les frontières bruit/parole, sans toutefois les identifier comme telles.
L'utilisation de cet algorithme seul permet d'isoler un vecteur d'observations représentatif de chaque segment comme entrée du programme de reconnaissance.

2.2. La détection automatique bruit/parole

Comme nous l'avons vu, le module de segmentation fournit une liste de frontières correspondant aux ruptures du signal. Cependant, on ne sait pas a priori pour chacun des segments s'il est situé dans une zone de parole ou dans une zone de bruit, et à plus forte raison quelles sont les frontières de début et fin de parole dans le signal. Bien entendu, la présence de bruit interdit de se limiter à un seuil fixe sur l'énergie pour discriminer les segments. Dans la littérature, de nombreux détecteurs bruit/parole sont proposés, mais rares sont ceux qui tirent parti d'un algorithme de segmentation.
Diverses techniques sont fondées sur la reconnaissance des formes, comme [Lamel 81], d'autres sur l'utilisation de coefficients acoustiques particuliers comme le taux de passage par zéro, l'énergie,... [Junqua 92].
L'approche que nous avons retenue est basée sur le fait que même si l'énergie du signal n'est pas un paramètre robuste en milieu bruité, ses maxima d'amplitude correspondent toujours à des noyaux vocaliques. Deux méthodes ont été étudiées, l'une basée sur l'analyse temporelle du signal, l'autre sur son analyse fréquentielle [Puel 94].
Leur fonctionnement est similaire : dans un premier temps, le paramètre extrait permet de localiser les noyaux vocaliques et d'opérer un premier étiquetage "parole". Dans un second temps une coordination temporelle avec les résultats de la segmentation fournit les résultats désirés.
Nous décrivons ici la méthode temporelle : l'étude de la variation de l'amplitude du signal temporel par l'intermédiaire de la mesure de son abscisse curviligne va nous permettre de discriminer les segments de parole des segments de bruit et de procéder au premier étiquetage. Ensuite, des règles de durée et de proximité, basées sur la connaissance que l'on a du signal et sur les résultats fournis par la segmentation, permettront de revenir sur cet étiquetage et de l'affiner.

L'étiquetage statique

Dans un premier temps, l'abscisse curviligne $s(t)$ du signal de parole $y(t)$, où t est l'indice des échantillons, est calculée. Soit la fonction :

$$S(n) = s(nL) - s((n-1)L)$$

où L est un nombre d'échantillons fixé (une trame).
$S(n)$ représente une valeur moyenne de la "longueur de la courbe" par unités de temps. En supposant que le bruit est stationnaire pour chaque segment, la fonction S varie peu dans les zones de bruit, croît très sensiblement avec l'apparition de la parole pour décroître lors de sa disparition. Les moyennes $\overline{S_i}$ et écart-types $\sigma(S_i)$ de S pour le segment i représentent notre indicateur de bruit ou parole. Deux seuils sont utilisés : λ_1 et λ_2, calculés automatiquement sur les trames du signal supposées n'être que du bruit (et où S présente ses minima).
L'étiquetage statique consiste à comparer les moyennes et écart-types de S par segment à ces seuils :

- les segments tels que $\overline{S_i} > \lambda_2$ sont étiquetés "parole",
- les segments tels que $\overline{S_i} < \lambda_1$ sont étiquetés "bruit",

- les segments tels que $\lambda_1 \leq \overline{S_i} \leq \lambda_2$ sont traités à l'étape suivante.

La coordination temporelle

Nous appliquons les règles suivantes à chacun des segments :

- un segment isolé de parole est classé "bruit".
- un court segment de bruit entre deux segments de parole sera classé "parole" s'il ne dépasse pas 80 ms (tenue de plosive).

Enfin, un traitement d'exceptions est prévu pour quelques situations particulières (SNR très faible, présence d'artefacts juste avant ou après le mot prononcé) où l'étiquetage statique se déroule anormalement. Typiquement, si moins de trois segments de parole sont détectés, les seuils sont réévalués et la détection est élargie autour du segment où l'activité "parole" est la plus vraisemblable. A l'issue de cette étape, nous disposons de l'étiquetage définitif des segments, et par là même, de la position de la parole dans le signal.

2.3. Le débruitage

Les techniques de débruitage du signal ont été largement étudiées en communication radio-mobile et en reconnaissance de mots isolés. Le point commun de la plupart de ces techniques est d'opérer la soustraction du bruit estimé dans le domaine spectral, permettant de résoudre pour une grande part le problème du bruit additif.
Nous avons implémenté l'algorithme de Soustraction Spectrale Linéaire du bruit (SSL) [Mokbel 92].
On considère que le signal $x(n)$ est l'addition de la parole $s(n)$ et d'un bruit aléatoire non corrélé $b(n)$, stationnaires à court terme.

$$x(n) = s(n) + b(n) \quad (1)$$

Dans le domaine spectral

$$\Gamma_x(\omega) = \Gamma_s(\omega) + \Gamma_b(\omega) \quad (2)$$

où $\Gamma_x(\omega)$ représente la densité spectrale de puissance à court terme de $x(n)$. La méthode générale de débruitage par soustraction spectrale consiste à calculer le signal débruité :

$$\hat{S}(\omega) = [|X(\omega)|^a - |\hat{B}(\omega)|^a]^{\frac{1}{a}} e^{j\phi_s(\omega)} \quad (3)$$

En reconnaissance de parole, la phase $\phi(\omega)$ apporte peu d'informations, aussi nous limitons nous à l'amplitude. Le paramètre a est fixé à 1 (soustraction spectrale en amplitude). On se limite donc à :

$$\hat{S}(\omega) = |X(\omega)| - |\hat{B}(\omega)| \quad (4)$$

Toute la difficulté réside dans l'estimation du bruit $\hat{B}(\omega)$. Nous utilisons alors les résultats du détecteur bruit/parole : la plage de bruit la plus longue détectée juste avant ou après le mot prononcé nous permet de mettre à jour l'estimation du bruit courant. Le spectre du bruit est estimé sur le plus grand nombre possible de fenêtres de 32 ms, avec un recouvrement de 16 ms. Le vecteur représentatif conservé est la moyenne du spectre sur toutes ces fenêtres. Il est possible de faire porter un coefficient multiplicateur sur l'énergie soustraite du bruit, pour accentuer l'effet du débruitage : coefficient constant ou fonction du SNR.

2.4. L'analyse spectrale

Après segmentation du signal et éventuellement détection bruit/parole et débruitage, un vecteur d'observations est calculé par segment, comprenant 8 coefficients MFCC, leurs 8 dérivées, l'énergie de la fenêtre, sa dérivée, et la longueur du segment.

2.5. Le système de reconnaissance

Le système de reconnaissance utilisé est basé sur les Modèles de Markov Cachés. L'ensemble du vocabulaire de l'application est décrit par réseaux de pseudo-diphones. Les vecteurs d'observation du systèmes sont constitués lors de l'analyse spectrale, et composés au moins des 8 coefficients MFCC et de la longueur du segment. Différentes dimensions du vecteur d'observations ont été testées, en utilisant tout ou partie des données disponibles. Les modèles acoustiques sont classiques et élémentaires, les lois d'observation sont des gaussiennes simples de matrice de covariance diagonale.

3. EXPÉRIMENTATION

Deux séries d'expérimentations ont été menées successivement,

- sur le corpus ARS : 43 mots isolés prononcés par 4 locuteurs dans un véhicule roulant à différentes vitesses,
- sur un corpus CNET : test d'un système de reconnaissance de 16 mots prononcés par une centaine de locuteurs au travers de plusieurs réseaux téléphoniques :
  - le réseau analogique appelé RTC,
  - le réseau numérique NUMERIS,
  - le réseau radio-téléphone, pour diverses conditions de roulement, appelé ITINERIS.

Nous avons essentiellement étudié les réseaux RTC et ITINERIS, les réseaux RTC et NUMERIS présentent des conditions très proches. Les résultats obtenus permettent de mesurer l'apport spécifique de chacun des modules.

3.1. Résultats sur le corpus ARS

Le tableau 1 montre les différences mesurées en millisecondes entre les résultats du module de détection bruit / parole et un étiquetage manuel du corpus.
Rappelons que l'étiquetage manuel du corpus intègre un relâchement systématique des frontières de l'ordre de 100 ms. On constate une différence systématique de 60 ms dans la détection des débuts de mots, certainement dûe au relâchement manuel trop important. Les fins de mots sont plus difficiles à observer, avec une différence moyenne de 150 ms (retard ou avance). Dans tous les cas, les différences rencontrées sont du même ordre d'idée que le relâchement manuel : le programme de détection automatique

	130 km/h	90 km/h	Total
Locuteur 1			
Début	65	60	63
Fin	109	195	152
Locuteur 2			
Début	49	52	51
Fin	97	206	152
Locuteur 3			
Début	53	56	55
Fin	111	250	181
Locuteur 4			
Début	78	91	85
Fin	91	128	110

TAB. 1 - : Différences entre segmentation automatique et manuelle

peut donc avantageusement remplacer un étiquetage manuel des frontières sur le corpus.

3.2. Résultats sur le corpus CNET

Les tableaux suivants montrent le taux d'erreur constaté lors d'expériences de reconnaissance sur les corpus CNET, selon le prétraitement utilisé. Les systèmes de reconnaissance évalués sont indépendants du locuteur, dans le sens où les ensembles d'apprentissage et de test sont disjoints quant aux locuteurs. Pour chaque système, nous réalisons quatre expériences qui dépendent des conditions d'enregistrement :

- apprentissage sur le réseau RTC et test sur le réseau RTC,
- apprentissage sur le réseau ITINERIS et test sur le réseau RTC,
- apprentissage sur le réseau RTC et test sur le réseau ITINERIS,
- apprentissage sur le réseau ITINERIS et test sur le réseau ITINERIS.

	Réseau RTC	Itineris
Apprentissage ITI	18.4 %	7.2 %
Apprentissage RTC	5.6 %	10.9 %

TAB. 2 - : Segmentation seule

Les résultats obtenus par le module de segmentation seul (tableau 2) nous servent de référence pour évaluer les autres modules. Au tableau 3, le modèle réalisé ne comprend pas de description du bruit avant et après chaque mot : il est donc très sensible aux petits écarts du détecteur bruit/parole. En entourant le modèle de chaque mot d'un modèle de bruit, (tableau 4), le système est plus robuste.

La soustraction spectrale du bruit apporte une nette amélioration au système, (tableau 5), principalement lorsqu'on

	Réseau RTC	Itineris
Apprentissage ITI	12.4 %	11.6 %
Apprentissage RTC	6.9 %	19.6 %

TAB. 3 - : Segmentation, détection bruit/parole

	Réseau RTC	Itineris
Apprentissage ITI	17.9 %	7.8 %
Apprentissage RTC	5.6 %	11.6 %
Apprentissage RTC-ITI	4.9 %	7.4 %

TAB. 4 - : Segmentation, détection bruit/parole

teste sur un environnement un système appris sur un autre environnement.

4. CONCLUSION

Le prétraitement présenté ici s'avère efficace, principalement dans le contexte qui nous intéresse : le changement d'environnement, où les erreurs de reconnaissance sont réduites de plus de 25 %. Les améliorations en cours de ce système concernent la modélisation du canal de communication dans le débruitage (ce n'est plus un bruit additif, mais convolutif), ainsi que le couplage de ce prétraitement segmental à une application de type centiseconde.

Références

[André-Obrecht 88] R. ANDRÉ-OBRECHT, "A New Statistical Approach for Automatic Segmentation of Continuous Speech Signals", IEEE Trans. on ASSP, vol. 36 pp 26-40, January 1988.

[Junqua 92] B. MAK, J.C. JUNQUA, B. REAVES, "A robust speech/non-speech detection algorithm using time and frequency-based features", ICASSP 1992.

[Lamel 81] L.F. LAMEL, L.R. RABINER, A.E. ROSENBERG, J.G. WILPON "An Improved Endpoint Detector for Isolated Word Recognition", IEEE Trans. on ASSP, Vol. 29, pp 777-785, August 1981.

[Mokbel 92] C. MOKBEL, "Reconnaissance Automatique de la Parole dans le Bruit", Thèse de Doctorat, TELECOM PARIS, juin 1992.

[Puel 94] J-B. PUEL, R. ANDRÉ-OBRECHT, "Robust Signal Preprocessing for HMM Speech Recognition in Adverse Conditions", ICSLP 1995.

	Réseau RTC	Itineris
Apprentissage ITI	11.63 %	7.57 %
Apprentissage RTC	4.46 %	8.74 %
Apprentissage RTC-ITI	4.19 %	7.77 %

TAB. 5 - : Segmentation, détection bruit/parole, SSL