METHODES DE SEGMENTATION SYLLABIQUE EN RECONNAISSANCE DE LA PAROLE

D. Fohr*, J.P. Haton*, F. Lonchamp**, L. Sauter***

*CRIN - ** Institut de Phonétique de Nancy - *** CGE Marcoussis

ABSTRACT

Segmentation at the acoustic-phonetic level is a major step in continuous speech recognition. We examine such processes in two different contexts : pattern matching based recognition and knowledge based recognition.

Segmentation into syllables seems to be an interesting step, considering the importance of the syllable in the phonation process. Also, the phonetic expert relies heavily on the syllable for his reasoning.

We present two methods for segmenting into syllable-based segments, developped independently in Marcoussis and Nancy. These methods are compared on common data.

INTRODUCTION

Les problèmes de segmentation au niveau acoustico-phonétique constituent un point majeur de la reconnaissance de parole continue. Nous envisageons ici ces processus comme une étape dans le décodage phonétique par des méthodes de comparaison de prototypes d'une part, et de raisonnement fondé sur une base de connaissance d'autre part.

La segmentation syllabique apparaît comme une étape intéressante compte tenu de l'importance de la syllabe dans le processus de phonation. De plus, l'expert phonéticien accorde également une place importante à la syllabe dans son raisonnement.

Nous allons présenter ici deux méthodes de segmentation syllabique développées indépendamment à Marcoussis et à Nancy, avec des finalités complémentaires. Ces méthodes ont été comparées sur des données communes.

SEGMENTATION ET COMPARAISON DE PROTOTYPES

Utilisation de la segmentation pour la comparaison de prototypes

Une des approches de la reconnaissance de la parole utilise des techniques de reconnaissance des formes : elle consiste à comparer des portions plus ou moins longues du signal de parole à des prototypes.

Si les portions sont très courtes, représentant un spectre à court terme, le résultat du traitement sera, par exemple, une suite d'étiquettes de noms de phonèmes. C'est le cas des méthodes dites centisecondes. Un phonème pourra être hypothésé lorsqu'une densité suffisante d'étiquettes identiques sera trouvée. La connaissance d'une segmentation du signal pourra donner des points d'ancrage pour l'émission d'hypothèses. Dans RAPACE [1], une segmentation en demi-syllabes permet d'éviter que plusieurs voyelles soient hypothésées pour un seul noyau vocalique. D'autre part, on tient compte du fait qu'en français, seules certaines suites de consommes peuvent initier ou terminer une demi-syllabe.

On utilise souvent une portion de signal correspondant à l'élocution d'un mot entier : il s'agit alors de la reconnaissance de mots isolés. La programmation dynamique permet de trouver un chemin de distorsion temporelle qui aligne le mot inconnu et les prototypes. Il est connu que cette technique est améliorée lorsqu'on ajoute des contraintes réalistes de pente pour le chemin de distorsion. La réduction de l'espace de recherche qui en résulte diminue à la fois la complexité des calculs et le nombre d'erreurs de reconnaissance en évitant qu'un chemin fantaisiste permette d'aligner deux mots différents.

Si les deux mots à comparer sont segmentés de façon analogue, alors l'espace de recherche peut encore être réduit en ajoutant comme contraintes d'aligner les points correspondants.

Si les segments sont peu sensibles au contexte, comme c'est le cas pour les syllabes et les demi-syllabes, alors il sera possible de remplacer les prototypes de mot par des concaténations de prototypes de segment.

Un certain nombre de systèmes de reconnaissance ont été simulés à partir des considérations précédentes. En ce qui concerne l'approche centiseconde, une segmentation en demi-syllabes a été utilisée dans RAPACE. En reconnaissance de mots isolés, le même algorithme de segmentation a permis d'utiliser un corpus de segments multilocuteurs comme ensemble de référence [2].

Nous décrivons maintenant succintement l'algorithme de segmentation tel qu'il a été utilisé dans ces deux cas :

Algorithme de segmentation

Notre algorithme de segmentation est fondé sur le principe suivant : tout maximum et tout minimum significatif de l'amplitude subjective du signal de parole est un candidat potentiel pour être une frontière entre deux demi-syllabes. Une mesure de l'amplitude subjective peut être obtenue en calculant l'énergie du signal de parole après un filtrage approprié.

Selon ce principe, la parole est segmentée en recherchant des minimums et des maximums successifs de l'énergie du signal. Ceux-ci correspondent respectivement aux groupes de consonnes et aux noyaux vocaliques. Afin d'éliminer des extremums non significatifs, nous utilisons la méthode suivante :

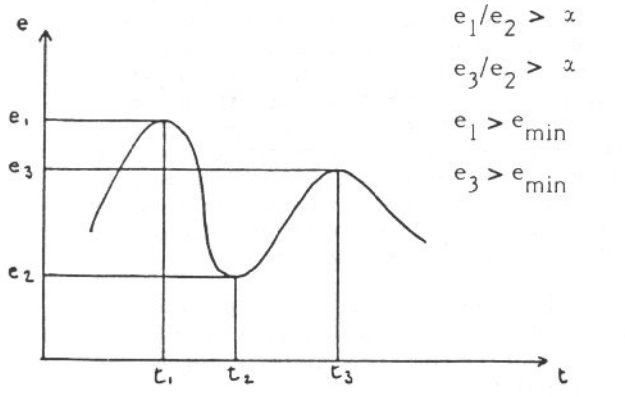
Une suite d'extremums de l'énergie est recherchée vérifiant les contraintes suivantes (cf fig. 1) :

- la longueur de chaque demi-syllabe doit dépasser un temps minimum (de l'ordre de 40 millisecondes)

- le rapport entre les valeurs de l'amplitude au début et à la fin de chaque demi-syllabe doit dépasser une certaine valeur.

- les maximums retenus doivent correspondre à une énergie suffisante.

Figure 1 : Segmentation



contraintes

$t_{i+1} - t_i > \ell_{min}$

$e_1/e_2 > \alpha$

$e_3/e_2 > \alpha$

$e_1 > e_{min}$

$e_3 > e_{min}$

La fonction utilisée pour mesurer l'énergie est obtenue après filtrage passe-haut du signal de parole (fréquence de coupure : 500 Hz), donnant ainsi une mesure correspondant à l'amplitude subjective. Cette fonction est mesurée toutes les 8 ms puis légèrement lissée.

Résultats

Cet algorithme a été évalué sur un corpus de 10 phrases équilibrées extraits de [3]. (Note : il ne s'agit pas du même corpus que ci-dessous. Des résultats sur le même corpus seront disponibles ultérieurement). Sur ces données, le nombre d'insertions est inférieur à 5 %, le nombre d'omissions est inférieur à 6 %.

SEGMENTATION ET RAISONNEMENT FONDE SUR UNE BASE DE CONNAISSANCE

Démarche de l'expert et fonctionnement général du système

Nous développons depuis deux ans un système pour le décodage acoustico-phonétique de la parole. Notre but est le décodage multilocuteur en parole continue. Pour ce faire, nous utilisons les connaissances d'un expert phonéticien (François Lonchamp) et nous formalisons son expertise pour le système expert SYSTEXP. François Lonchamp est capable, à partir d'une représentation visuelle du signal acoustique de parole (spectrogrammes), de décoder des phrases prononcées de manière naturelle avec un taux de reconnaissance nettement supérieur à celui des algorithmes actuellement disponibles [4].

Quand l'expert phonéticien décode un spectrogramme, il commence par jeter un regard sur l'ensemble du spectrogramme pour calculer la durée moyenne vocalique. Cette notion lui est indispensable pour segmenter les paquets vocaliques (plusieurs voyelles et sonnantes sans frontières marquées) ou les zones qui peuvent être segmentées de plusieurs manières (consonne ou voyelle très longue).

D'autre part, l'expert ne semble pas avoir de problème de segmentation dans les cas non ambigus : il y a séparation entre segmentation et étiquetage : segmentation grâce à des frontières nettes puis étiquetage phonétique en appliquant des règles contextuelles en fonction d'indices pertinents. Dans les cas difficiles (paquets vocaliques) segmentation et identification vont de pair, avec essai de différentes possibilités et choix de celle qui semble la plus probable.

Le système SYSTEXP se fonde sur la démarche de l'expert humain. Dans une première phase, nous effectuons un prétraitement pour obtenir la durée vocalique moyenne et une segmentation grossière en grandes classes phonétiques. Puis, dans une deuxième phase, nous utiliserons un moteur d'inférence et les règles obtenues avec l'expert pour identifier et segmenter finement. Les règles peuvent appeler les procédures de traitement du signal pour extraire les indices nécessaires pour prendre une décision. De plus les règles peuvent modifier la segmentation, c'est-à-dire changer les limites d'un segment, scinder un segment en deux ou regrouper deux segments. La première phase est implantée sous forme procédurale alors que la deuxième est sous forme d'un système expert (cf figure 2).

	Nom	Résultats	Implantation
(1)	NOVOCA	durée vocalique moyenne noyaux vocaliques + limites	procédural
	PLOSI	plosives + limites	procédural
	FRICA	fricatives + limites	procédural
(2)	EXP	phonèmes frontières précises	système expert

(1) Prétraitement
(2) Identification et segmentation

Figure 2 : Synoptique du système SYSTEXP

Fonctionnement du système de segmentation NOVOCA

Le but de Novoca est de trouver tous les noyaux vocaliques contenus dans une phrase et de déterminer la durée moyenne vocalique. L'expert nous a tout d'abord décrit les critères qu'il utilisait pour reconnaître les noyaux vocaliques sur un spectrogramme : forte intensité, présence de formants (pics d'énergie dans le spectre) dans une certaine bande de fréquence ... Nous avons décidé d'utiliser un critère sur l'énergie. Après avoir numérisé (10 kHz 10 bits) et préaccentué le signal de parole, nous calculons tout d'abord l'énergie dans une bande de fréquence. Cette bande a été choisie de manière à défavoriser les sons ayant principalement de l'énergie en très basse fréquence (par exemple les nasales) et ceux qui ont de l'énergie en haute fréquence (par exemple les fricatives). Il fallait inclure dans cette bande la zone du premier et du deuxième formant des voyelles. Nous avons essayé différentes fréquences (250-1500 Hz, 250-2230 Hz, 250-2500 Hz) et la bande retenue a été la bande 250-2350 Hz. Puis nous cherchons les pics de cette courbe qui vérifient les critères ci-après :

- le nouveau pic doit atteindre au moins 55 % du pic précédent (deux noyaux successifs ne peuvent pas avoir des énergies trop différentes),
- la vallée de part et d'autre du pic est fonction de la hauteur du pic (plus un pic est important, plus la vallée doit être importante),
- au moins 50 % des échantillons du noyau vocalique doivent être voisés.

Quand un pic vérifie tous ces critères, on recherche le début et la fin du noyau correspondant. On ne décrira que la recherche de la fin du noyau car la recherche du début est symétrique.

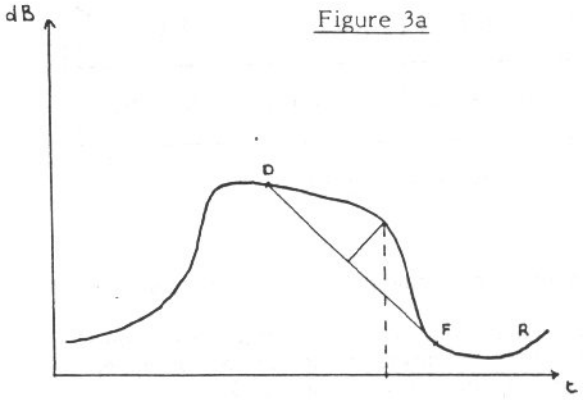
(1) A partir du pic, on recherche le point D où l'énergie est inférieure ou égale à E(PIC)-SEUIL1 (point de début de la descente).

(2) On recherche ensuite le point R où l'énergie commence à remonter.

(3) A partir de R, on recherche le point F qui vérifie E(F) > E(R)+SEUIL2

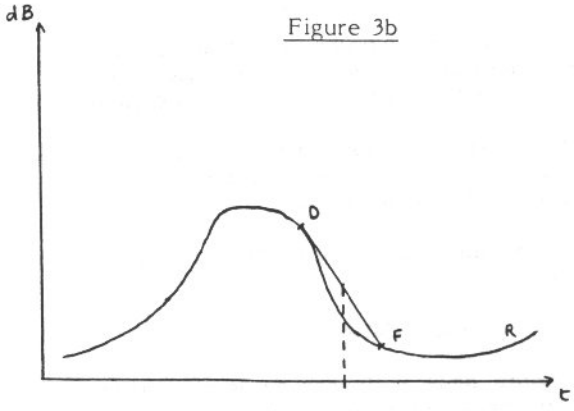
(4) On trace le segment D,F et on cherche le point de la courbe de l'énergie situé au-dessus de cette droite et qui est à la plus grande distance de cette droite. Si le point trouvé est entre D+(F-D)/4 et F-(F-D)/4 c'est le marqueur de fin de noyau sinon c'est (D+F)/2. La figure 3 présente deux exemples.

Figure 3a



Exemple d'une courbe d'énergie présentant une épaule

Figure 3b



Exemple sans épaule

On effectue 2 passes :

une première passe en imposant une vallée importante de part et d'autre du pic (pics ayant une très forte probabilité d'être des noyaux mais taux d'omission de noyaux important).

une deuxième passe en imposant une faible vallée de part et d'autre du pic (taux d'omission très faible mais quelques insertions).

Résultats

Le corpus est constitué de 57 phrases équilibrées de Combescure [3], prononcées à un rythme naturel d'élocution par cinq locuteurs masculins non professionnels. Les phrases étaient lues par un locuteur expérimenté et répétées de mémoire. Le rythme et la prosodie étaient imposés (> 14 phonèmes par seconde). Le corpus comprenait 581 noyaux vocaliques, les résultats sont donnés figure 4.

	Trouvés	Insertion	Omissions
1ère passe	563 (97 %)	18 (3 %)	18 (3 %)
2ème passe	572 (98 %)	26 (< 5%)	9 (< 2%)

Figure 4 : Résultats

Les erreurs sont dues principalement à 5 raisons :

- omission de noyaux faibles situés à côté de consonnes très intenses et insertion de cette consonne, (dans 'toujours' /ʒ/ plus intense que /u/)
- omission de /i/ presque complètement assourdi, en tête ou en queue d'énoncé,
- omission de phonèmes trop brefs (rythme très rapide),
- découpage de voyelles nasales en deux parties (importante fluctuation d'énergie au cours de la voyelle). Ceci est compté comme une insertion,
- insertion d'un noyau supplémentaire dans les logatomes /bR/ /dR/ (remontée d'énergie après la plosive puis chute dans le /R/).
- insertion de quelques consonnes (/l/ /m/ /n/) très vocaliques.

CONCLUSION

Nous avons présenté deux méthodes de segmentation syllabique développées indépendamment avec des finalités différentes. Nous pensons qu'il serait intéressant d'examiner dans quelle mesure leur utilisation conjointe pourrait permettre d'obtenir une segmentation syllabique de qualité encore meilleure.

REFERENCES

[1] L. Sauter, "RAPACE : un système de reconnaissance analytique de parole continue", 4 ème congrès AFCET RFIA, Paris, janv. 1984.

[2] L. Sauter, "Speaker independent isolated word recognition using a segmental approach", Proc. ICASSP 85, Tampa, mars 1985.

[3] P. Combescure, 20 listes de phrases phonétiquement équilibrées, Rev. d'Acoust. n° 56, pp. 34-42, 1981.

[4] N. Carbonnel, D. Fohr, J.P. Haton, F. Lonchamp, J.M. Pierrel, "An Expert System for the Automatic Reading of french Spectrograms", Proc. ICASSP 84, San Diego, mars 1984.