SEGMENTATION ET RECONNAISSANCE EN PAROLE CONTINUE A L'AIDE DES REFERENCES ISSUES DU SYSTEME VARAP.

RINGOT P. ANDREEWSKY M. DEVILLERS L. DESI M. PARISSE C.

CNRS - LIMSI - ORSAY BP 30 91406 ORSAY CEDEX FRANCE

We present two possible approaches of continuous speech recognition. The first uses a segmentation obtained by a training process. The second using an appropriate distance allows to simultaneously achieve the segmentation and recognition.

Dans ce travail, on expose les expériences de reconnaissance qui ont été faites, compte tenu du système d'étiquetage automatique employé (utilisé sur un corpus de 700 phrases) et du mode de sélection utilisé dans le système VARAP. Deux méthodologies différentes sont exposées. L'une qui procède d'abord à une segmentation, puis à une reconnaissance, la seconde effectue ces deux opérations simultanément. Dans ce qui suit, on utilise une distance qui est donnée par la formule :

$$| (O1 - X1) - (O2 - X2) | + \ldots + | (O15 - X15) - (O16 - X16) |$$

où les valeurs O1...O16 sont les 16 valeurs du premier spectre et X1...X16 sont les valeurs du second.

I. SEGMENTATION OBTENUE A PARTIR DU CORPUS D'APPRENTISSAGE.

Quatre paramètres sont définis: Em, Tm, EM, TM dont la signification est la suivante:
Em est un seuil minimal d'énergie.
Tm est un seuil minimal temporel.
EM est un seuil maximal d'énergie.
TM est un seuil maximal temporel.

Le mode d'utilisation de ces quatre paramètres est le suivant :
Toutes les alternances qui ont une différence d'énergie inférieure à Em sont lissées tant que leurs fluctuations sur le temps ne sont pas supérieures à TM. De même, toutes les alternances qui ont une fluctuation sur le temps inférieure à Tm sont lissées tant que les fluctuations sur l'énergie ne sont pas supérieures à EM. Les paramètres Em, EM, Tm, TM sont déterminés sur le corpus d'apprentissage. Pour cela, on sélectionne le plus petit écart non lissé sur l'énergie et sur le temps dans chaque phrase étiquetée du corpus et on calcule pour chacun de ces écarts le nombre de fois dans le corpus où il a été conservé ou lissé Le plus grand des plus petits écarts de chaque phrase du corpus donne les valeurs de EM et de TM.
La sélection des seuils inférieurs Em et Tm se fait en imposant un rapport aussi optimisé que possible entre le nombre de fois où Em, Tm ont été lissés et le nombre de fois où ils ont été conservés, étant entendu que l'optimisation est définie par le plus petit pourcentage possible d'erreurs conservées sur le corpus.
Les résultats de la segmentation sur 50 phrases après optimisation sont les suivants :
Dans les 50 phrases, il y a en tout 1032 segments.
Le lissage optimisé laisse un nombre total d'ajouts égal à 29 et un nombre total d'élisions égal à 28.
Parmi les ajouts, il y a 14 segments qui correspondent ou bien à des répétitions du même phonème (par exemple /sss/), ou bien à des dissociations d'étiquettes complexes du type voyelle-voyelle (par exemple /aa/ qui donne /a/,/a/) ou bien des étiquettes complexes du type voyelle-liquide (par exemple /ar/ qui donne /a/,/r/).

Il reste donc 15 ajouts qui ne peuvent pas être interprétés dans le cadre du niveau phonétique où nous nous situons.
Pour les élisions, 18 segments correspondent à des étiquettes complexes que la segmentation n'a pas dissociées. Ce sont des groupes du type consonne-liquide ou consonne-chuintante (par exemple /b/ et /r/ formant une seule étiquette /br/) ou encore des groupes du type consonne-consonne (par exemple /t/ et /d/ ne formant qu'une seule étiquette /td/).
Il reste donc 10 segments d'élisions.
Par conséquent, le pourcentage total d'erreurs est de l'ordre de 5% . Si on se refère à une ségmentation phonémique, le pourcentage d'erreurs est de 3% dans notre système de référence qui admet des étiquettes phonétiques multiples.

II. UNE EXPERIENCE DE SEGMENTATION ET DE RECONNAISSANCE SIMULTANEE.

Cette expérience comprend les étapes suivantes:

A. Sur la suite continue des spectres correspondants à un énoncé donné, on prélève les triplets successifs de spectres en commençant dans l'ordre par le premier spectre puis le second et ainsi de suite. Deux triplets consécutifs ont donc toujours deux spectres en commun.

B. Ces triplets sont proposés au dictionnaire des références ternaires obtenues à partir du système VARAP. On obtient ainsi des treillis de quatre phonèmes candidats, résultats d'un scrutin majoritaire effectué sur les quinze plus proches références du dictionnaire où l'on tient compte de la position des références et de leur nombre.
Simultanément, on conserve la distance entre le meilleur candidat du treillis et la référence analysée, et on affiche la courbe des distances.

C. Résultats.
Un dépouillement effectué sur 200 phrases montre tout d'abord que la courbe des distances suit les contrastes de la courbe d'énergie et qu'elle fournit une segmentation de même qualité.

Cela nous conduit à faire la remarque importante suivante: l'analyse centiseconde effectuée utilise un dictionnaire de références qui ne contient pas les transitions. Par conséquent, on pouvait s'attendre à ce que les extrema de la coube d'énergie soient aux minima de la courbe des distances. Une explication de ce phénomène tient dans le fait que la distance utilisée bien qu'étant du type convergence uniforme contient un facteur d'énergie qui se manifeste de manière importante sur les voyelles.

Les comparaisons avec le corpus d'apprentissage font apparaître:

- des plages de grande stabilité phonétique permettant de déterminer des îlots de confiance.
- la possibilité en cours d'élaboration d'identifier l'énoncé à partir de la suite des treillis et de la disposition des extrema de la courbe des distances à l'exception des débuts et des fins d'énoncé.

Des dépouillements effectués sur 200 phrases donnent des résultats de reconnaissance phonétique très variables selon les phrases, de l'ordre de 60 à 75 % .
Les procédures indispensables permettant de dégager une décision à partir des treillis et des extrema sont en cours d'élaboration.

BIBLIOGRAPHIE

DESI M. POIRIER F. "Le système SHERPA: étiquetage et classification automatique par apprentissage pour le décodage automatique de la parole continue", Thèse en Sciences, Paris-Sud, Orsay, 1985.

LAZREK M. HATON J.P. "Segmentation et identification des phonèmes dans un sysème de reconnaissance automatique de la parole continue", Congrès AFCET Reconnaissance des Formes et Intelligence Artificielle. Paris, Janvier 1984, p. 5.

MARIANI J. "ESOPE: un système de compréhension de la parole continue", Thèse d'Etat, Université Paris VI, 9 juillet 1982.

MARIANI J. "Méthodes en reconnaissance phonétique", 11ème ICA. Toulouse, 125-137, 1983.

MERCIER G. GERARD M. GILLET D. NOUHEN-BELLEC A. QUINTON P. SIROUX J. "Le système de reconnaissance de la parole continue KEAL", 12ème JEP, GALF, Montréal, 1981.

MERCIER G. "Analyse acoustique et phonétique dans le système KEAL", 12ème JEP, GALF, Montréal, 1981.