JOURNAL DE PHYSIQUE IV
Colloque C5, supplément au Journal de Physique III, Volume 4, mai 1994

Apport d'une segmentation statistique du signal de parole dans un système de décodage acoustico phonétique basé sur les connaissances

R. ANDRE-OBRECHT et N. PARLANGEAU

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

ABSTRACT:

In a speech labelling system, traditionnaly the first stage consists in the transcription of the continuous acoustic signal into a series of discrete elementaries units. This stage called Acoustic-Phonetic decoding unrolls three phasis : parametrisation, segmentation and identification.

All the systems do not use the module of segmentation stricktly speaking. For those that use it, they differ from one another in their strategy of segmentation.So we distinguish two different approachs, one is based on an a posteriori segmentation and the other on an a priori one.

This paper presents an automatic phoneme recognition system based on a preprocessing segmentation algorithm called the ***Divergence*** *method. The study of these performances got by Divergence and those obtained by a knowledge based system (SED) where the segmentation is an a posteriori result, has brought us to propose an hybrid system named the SED-Divergence method, using the advantages of the two methods.We asser these systems by integrating them into a labelling process.*

1. INTRODUCTION

Dans un système de reconnaissance automatique de la parole, comme dans un système d'étiquetage automatique, traditionnellement la première étape consiste à transcrire le signal acoustique de nature continue en une suite discrète d'unités élémentaires (phonème...). Cette étape appelée Décodage Acoustico-Phonétique (D.A.P) se déroule généralement en trois phases : la paramétrisation du signal, la segmentation et l'identification.

On trouvera des méthodes différentes pour chaque phase. La paramétrisation peut être temporelle ou être issue d'une transformation du signal dans un nouvel espace; tel que l'espace temps-fréquence. L'identification peut faire appel à des méthodes de reconnaissance des formes statistique (quantification vectorielle, modèles de Markov cachés...)[1], de reconnaissance des formes structurelle[2] ou encore à base de connaissance [3].

Tous les systèmes ne mettent pas en oeuvre un réel processus de segmentation et bien souvent segmentation et identification sont confondues. Cependant, pour les systèmes utilisant la segmentation, on distingue de façon schématique deux stratégies de segmentation :

- ☞ une partition de l'espace des paramètres permet d'identifier des trames de 10 ms ; chaque changement d'étiquette motive une nouvelle frontière. C'est une **segmentation a posteriori.**
- ☞ une analyse statistique permet par l'intermédiaire d'une modélisation du signal de confronter deux hypothèses ; la décision sera interprétée comme une absence ou une présence de rupture. C'est une **segmentation a priori.**

Afin d'évaluer ces deux approches, nous nous sommes intéressés au système Segmentation Etiquetage Discontinuités (S.E.D) basé sur la première approche[4] et nous avons conçu le système Divergence qui tient de la seconde approche. A l'issue de cette étude, nous proposons un système hybride SED-Divergence qui exploite les avantages des deux méthodes et en pallie les inconvénients.

Article published online by EDP Sciences and available at http://dx.doi.org/10.1051/jp4:19945100

2 . LA METHODE S.E.D

La méthode S.E.D est basée sur une segmentation a posteriori du signal de parole. Après avoir procédé à une analyse temporelle du signal, un ensemble de règles permet d'étiqueter chaque trame du signal et d'en effectuer un regroupement afin d'obtenir la segmentation et l'étiquetage finals [5].

2.1 Paramétrisation

Les paramètres calculés sont des paramètres d'ordre temporel : l'amplitude, le taux de passages à zéro ainsi que l'énergie contextuelle et relative.
La méthode se décompose en trois grandes phases : un prétraitement dont le but est d'adapter les conditions d'enregistrement, une micro-modélisation non-linéaire et une optimisation .
La micro-modélisation est basée sur des transformations morphologiques qui vont permettre d'amplifier des discontinuités pertinentes du signal . L'optimisation consiste à trouver un modèle optimal d'approximation au sens des moindres carrés sur une fenêtre. Son application est un ultime lissage.
Tous les paramètres sont ensuite normalisés entre [0,1]. On utilise les dérivées premières et secondes de chacun des paramètres pour mesurer l'instabilité du signal.

2.2 Etiquetage en événements phonétiques

Sur chaque trame, on calcule un indice d'instabilité à partir des paramètres précédemment calculés ainsi qu'un indice phonétique pour chaque étiquette. L'étiquette ayant l'indice le plus fort est associée à la trame. L'ensemble des étiquettes utilisées est le suivant:

Etiquettes	Signification
K	Voyelle
M	Voyelle grave
N	Voyelle aiguë
L	Vocalique grave
U	Vocalique aiguë
O	Occlusif voisé
X	Occlusif sourd
F	Fricatif faible
Z	Fricatif voisé
S	Fricatif sourd
A	Attaque discontinue
B	Coda discontinu

figure 1 : étiquettes des événements phonétiques`

2.3 Identification et segmentation

La segmentation est motivée par l'identification des trames ; elle se fait par regroupement des étiquettes à partir d'un ensemble de règles acoustiques.Chaque nouvelle étiquette place une nouvelle frontière.

3 . LA METHODE DIVERGENCE

La méthode Divergence est un système basé sur une segmentation statistique a priori; chaque segment sera identifié par une quantification vectorielle à partir d'une paramétrisation d'ordre quéfrentiel.

3.2 Le prétraitement acoustique

La spécificité de notre approche réside dans l'utilisation de la segmentation comme prétraitement. La méthode utilisée est basée sur un test de décision statistique, la méthode de divergence Forward-Backward (F.B). Pour de plus amples détails relatifs à cette segmentation se référer à [6] .

Sur chaque segment sont ensuite calculés les huit premiers coefficients cepstraux ainsi que l'énergie (logiciel CNET) . L'extraction des paramètres se fait sans connaissance a priori de la nature des segments. Les paramètres sont estimés sur une fenêtre de 512 points centrée sur le milieu du segment (~30ms).

3.3 L'identification

L'étiquetage en classes phonétiques adopté est basé sur les principales caractéristiques des phonèmes, distinguant ainsi neuf grandes catégories (cf fig.2). L'avantage d'un tel ensemble d'étiquettes est de pouvoir caractériser des segments de nature infra-phonémique.

Etiquettes	Signification
K	Noyau vocalique central
M	Noyau vocalique grave
N	Noyau vocalique aigu
L	occlusive/liquide/nasale hautes freq.
U	occlusive/liquide/nasale basses freq.
O	Barre de voisement
Z	Fricative voisée
S	Fricative non voisée
X	Silence bruité
Q	Silence

figure 2 : étiquetage en classes phonétiques - Divergence -

L'identification se fait par quantification vectorielle non hiérarchique, c'est-à-dire qu'elle mène à une seule partition des données en classes: l'ensemble d'apprentissage est décomposé en nuages de points disjoints caractérisés par un point appelé centroïde. L'ensemble formé autour des centroïdes est appelé **dictionnaire.**
Nous construisons un dictionnaire pour chaque classe phonétique. Par la suite, l'identification de tout nouveau vecteur se fait par comparaison avec les représentants de tous les dictionnaires. La décision est prise au moyen de la règle du plus proche voisin [1].

4 . LA METHODE S.E.D-DIVERGENCE

Cette méthode est l'alliance entre la méthode S.E.D et la méthode Divergence. Le point de départ de cette alliance est la volonté d'allier les avantages respectifs des deux méthodes:

- ☞ l'identification des segments est faite de façon plus fine dans la méthode S.E.D que par quantification vectorielle,
- ☞ le taux de sur-segmentation de la méthode S.E.D est supérieur à celui de divergence F.B et de plus, les frontières localisées par la méthode de divergence F.B sont plus précises.

Nous avons donc conçu la méthode suivante: après segmentation du signal par la méthode Forward-Backward, on identifie chaque segment à partir de l'étiquetage trame par trame fourni par S.E.D. Une étiquette est affectée à un segment par un simple vote majoritaire sur les trames du segment.

5 . EVALUATION ET EXPERIMENTATIONS

Afin d'évaluer nos méthodes, nous avons procédé à un alignement automatique de la transcription phonétique à l'aide des étiquettes obtenu sur le signal, et ce à partir de modèles de Markov cachés compilés [7]. Nous avons ainsi défini un modèle par phonème, ce qui fait 40 modèles (cf. fig.3). Les traits en gras modélisent les transitions vides.

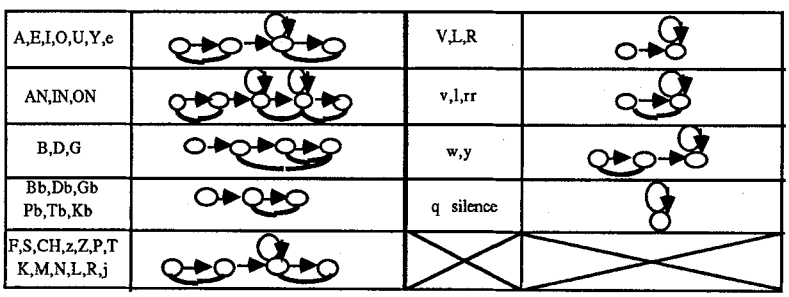


figure 3 : tableau des modèles de Markov

L'alignement se fait de la façon suivante : à partir de la phrase manuellement étiquetée en phonèmes, nous construisons un réseau global en tenant compte des coarticulations possibles. Les observations sont ensuite présentées à ce réseau.
L'apprentissage des modèles a été fait à partir des observations obtenues pour chaque phonème sur les 13 premières phrases d'un locuteur de la base EUROM0. La reconnaissance s'effectue sur les dix dernières phrases du même locuteur.

Nous avons mené deux expériences. Une première expérience a consisté à travailler avec des phrases dont les pauses ont été projetées sur le signal; chaque phrase est donc découpée en "sous-phrases". Dans la seconde expérience, les pauses ne sont pas projetées. Il apparait évident que les pauses sont une information importante pour éviter tout dérapage lors de l'alignement des observations : plus le réseau est long, plus le risque est grand. Il faut cependant remarquer que la non prise en compte des pauses pénalise davantage la méthode SED-Divergence. En effet, on voit que le taux d'erreurs augmente de 9% pour la méthode SED-Divergence entre les expérimentations avec ou sans pause, contre une augmentation de 3% pour la méthode Divergence.

	PAUSE			SANS PAUSE		
	reco	erreurs	omissions	reco	erreurs	omissions
Divergence	73 %	24%	2%	71%	27%	2%
SED - Divergence	73%	22%	5%	67%	31%	2%

figure 4 : tableau des résultats (erreurs à moins de 20 ms)

6 . CONCLUSION

Les résultats quantitatifs ne nous permettent pas à l'heure actuelle de tirer de conclusions fermes sur la question qui soutend ces travaux, à savoir quel est l'apport de la segmentation a priori dans un système à base de connaissances.
Les résultats ne paraissent pas excellents alors que le taux de frontières correctement trouvées par la méthode Divergence à moins de 20ms est de plus de 95 % ; ce taux est obtenu sans apprentissage de paramètres sur ce corpus.Une des raisons majeure est le faible ensemble de données utilisées aussi bien en phase d'apprentissage qu'en phase de test. Cependant, les résultats qualitatifs nous laissent penser que le segment est une information très intéressante. Pour approfondir cette voie, nous allons construire un système d'alignement automatique basé sur une segmentation a priori et sur des modèles de Markov cachés continus: les erreurs dûes à l'identification en événements phonétiques ou à la quantification vectorielle disparaîtront.
De plus, la méthode SED-Divergence permet de réduire les divers lissages appliqués lors du regroupement des étiquettes; une expérience intéressante consisterait à éliminer les lissages lors du calcul des paramètres temporels eux-mêmes, en introduisant l'information a priori contenue dans la segmentation basée sur la méthode de divergence Forward-Backward.

BIBLIOGRAPHIE

[1] H.Y SU , "Reconnaissance acoustico-phonétique en parole continue par Quantification Vectorielle: adaptation au locuteur .", Thèse de l'université de Rennes I . 1987.
[2] DALSGAARD P., BARRY W. , 3 Acoustic Phonetic features in the framework of Neural Network Multi Lingual Label Alignment", I.C.S.L.P., Kobe, Japan, Novembre 1990.
[3] EDERVEEN D., BOVE L. , " Knowledge based phoneme recognition ", Eurospeech volume II 1991.
[4] ANDRE-OBRECHT R., PERENNOU G., VIGOUROUX N., " Deux approches de l'étiquetage en événements phonétiques", 19èmes Journées d'Etudes de la Parole, mai 1992.
[5] KABRE H., "D.A.P. multilingue : système à base de connaissances et étiquetage automatique de corpus de parole ", Thèse de 3ème cycle de l'U.P.S Toulouse III, 1990.
[6] ANDRE-OBRECHT R., "A new statistical approach for automatic segmentation of speech signals", I.E.E.E trans. on ASSP1988.
[7] ANDRE-OBRECHT R., JACOB B.,De CALMES M., " Reconnaissance Automatique de la Parole dans le cadre de tres grands vocabulaires", 3ème Congrès Francais d'Acoustique Toulouse 1993.