Une approche de la détection des noyaux syllabiques et son utilisation en étiquetage phonétique automatique

H. Kabré, Guy Pérennou, Nadine Vigouroux

COLLOQUE DE PHYSIQUE
Colloque C2, supplément au n°2, Tome 51, Février 1990

1er Congrès Français d'Acoustique 1990

UNE APPROCHE DE LA DÉTECTION DES NOYAUX SYLLABIQUES ET SON UTILISATION EN ÉTIQUETAGE PHONÉTIQUE AUTOMATIQUE

H. KABRÉ, G. PÉRENNOU et N. VIGOUROUX

Laboratoire IRIT-CERFIA UA 824 CNRS, Université Paul Sabatier, 118, Route de Narbonne, F-31062 Toulouse Cedex, France

Résumé - Nous décrivons une approche de la syllabation par une méthode ascendante fondée sur une segmentation préalable en unités phonétiques. Après avoir donné les résultats obtenus sur des phrases de parole continue, nous examinons la contribution de la prise en compte de la structure syllabique dans les tâches d'alignement automatique effectuées par le systèmeVERIPHONE. Sa tâche consiste à aligner une transcription phonétique fournie par un expert phonéticien sur la chaîne d'événements phonétiques produite par le système SAPHO de manière purement ascendante.

Abstract - A description of an approach to parsing for syllables, through a bottom-up method that is based on prior segmentation into phonetic units. Results, secured from applying the method to phrases embedded in continuous speech, are given first. Next, the effect of taking into account syllabic structure is scrutinized, as this applies to the automatic alignment tasks (performed by VERIPHONE) involved when attempting to align a phonetic transcription (provided by an expert phonetician) onto the phonetic event string which is yielded by a strictly bottom-up labelling method (SAPHO System).

1 - INTRODUCTION

La syllabe joue-t-elle un rôle en tant qu'unité de décision dans le décodage phonétique de la parole ?
En psycholinguistique, des expériences comme celles de Mehler et coll. /1/ suggèrent qu'un traitement de niveau syllabique précède la reconnaissance de cibles phonémiques, mais la phonologie classique, en particulier générativiste, n'accorde qu'un rôle secondaire à la syllabe. Ce n'est que récemment qu'avec la phonologie métrique et la phonologie autosegmentale qu'elle reprend une place de premier plan. Mais son statut est encore loin d'être clair.
En reconnaissance automatique de la parole certains ont tenté d'introduire la syllabe dans le décodage phonétique (voir par exemple:/2/, /3/, parfois en liaison avec la prosodie /4/.
Actuellement plusieurs systèmes de reconnaissance de la parole, pour mieux modéliser la coarticulation, prennent pour modèles phonétiques de base des unités syllabiques —on peut citer comme exemple des systèmes basés sur la comparaison dynamique voir, /5/, /6/.
Un examen plus approfondi de ces systèmes montrerait que généralement les unités syllabiques y sont utilisées pour représenter les structures phonétiques et non comme unités de décision —dans de tels systèmes les décisions portent sur des unités plus grandes, mots ou groupes de mots, sans qu'il y ait de décisions intermédiaires partielles—.
A noter aussi que l'on a proposé des machines pour la reconnaisance de textes lus en détachant les syllabes les unes des autres. Celles-ci sont alors reconnues comme des unités isolées. Elles constituent dans ce cas les unités de décision /7/—voir aussi /8/ pour la reconnaissance du japonais.
Ainsi le statut des unités syllabiques en reconnaissance de la parole n'est pas simple. On manque d'ailleurs de bases objectives pour en apprécier l'apport. Notre communication vise à décrire une approche du décodage syllabique ascendant utilisant une segmentation préalable en événements phonétiques. Ce travail se fait en liaison avec un autre visant à développer un système d'étiquetage automatique de corpus de parole.
Dans l'environnement des bases de données acoustico-phonétiques, l'étiquetage automatique joue un rôle important /9/. L'un des objectifs du projet ESPRIT SAM /10/ est de développer de tels outils.
La contribution du traitement de niveau syllabique est évaluée sur le corpus EUROM-0 de SAM. Le critère choisi pour cette évaluation sera le gain de performances dans les tâches d'alignement automatique (effectuées par notre système VERIPHONE /11/) des transcriptions phonétiques sur le signal segmenté et étiqueté automatiquement en événements phonétiques (ceci est effectué par notre système SAPHO).

2 - DETECTION AUTOMATIQUE DES NOYAUX SYLLABIQUES

2.1 - Indices phonétiques normalisés

Le signal de parole digitalisé à 16Khz est analysé par un banc de 24 filtres dont la période d'intégration est fixée à 128 échantillons de signal avec recouvrement tous les 4 ms, soit 8 ms, que nous appellerons échantillon centiseconde ou plus brièvement frame.
Les paramètres retenus sont :

- énergie totale moyenne sur 7 frames,
- pente moyenne de l'énergie sur 7 frames,
- convexité moyenne de l'énergie sur 7 frames,
- un indice d'acuité élaboré sur la base des énergies dans les aigus,
- un indice dual traduisant la compacité du spectre.

Nous en dérivons un vecteur d'indices caractéristiques des classes d'événements phonétiques que nous avons retenus dans notre application d'étiquetage automatique. Ces indices sont i-noyau, i-noyauI, i-noyauU, i-liquide, i-friction, i-occlusion, i-explosion, i-non-parole. Tous ces paramètres et ces indices sont normalisés ce qui permet de s'affranchir des conditions particulières d'enregistrement et des caractéristiques du locuteur.

2.2 - Etiquetage des frames

L'étiquette d'un frame est déterminé par l'indice dont la valeur dépasse celle des autres indices : ainsi par exemple l'étiquette d'un frame sera K (initiale de Kernel) si la valeur de l'indice i-noyau dépasse celle des autres indices.
Les frames isolés (c.-à-d.: dont l'étiquette est différente de celle du frame précédent et de celle du frame suivant) sont réétiquetés. Voici des exemples de règles utilisées :

y —> x /x—x (si les deux frames adjacents portent l'étiquette x le frame y sera aussi réétiquetté x)
L —> K /K—O (un frame liquide noté L après un noyau K et avant une occlusive voisée O est converti en noyau K).
...

2.3 - Evénements phonétiques

Le tableau 1 représente la liste des événements phonétiques déterminés après l'étiquetage des frames. Ces événements constituent la chaîne d'entrée du système VERIPHONE.

SAPHO	VERIPHONE	Signification	SAPHO	VERIPHONE	Signification
V	K	Noyau	F	F	Fricative
VF	k	Noyau faible	FA	s	Etablissement de F
VA	y	Etablissement Noyau	FD	z	Coda de F
VD	v	Coda Noyau	RG	R	Liquide uvulaire
I	I	Noyau I	RA	b	Etablissement de R
IF	i	Noyau I faible	RD	c	Coda de R
IA	j	Etablissement de I	RX	X	R fricatif
ID	l	Coda de I	XA	h	Etablissement de X
U	U	Noyau U	XD	q	Coda de X
UF	u	Noyau U faible	NL	N	Nasale
UA	e	Etablissement de U	LJ	L	Liquide
UD	~	Coda de U	LA	a	Etablissement de N ou L
Q	Q	Occlusive sourde	LD	d	Coda de N ou de L
QD	>	Implosion de Q	US	*	Indéterminé
EX	<	Explosion de Q	NS	£	Non parole
O	O	Occlusive voisée			
OD	]	Implosion de O			
EV	[	Explosion de O			

Tableau 1 - Signification des événements phonétiques utilisés dans SAPHO et VERIPHONE.

Les segments composés de frames contigus de même étiquette constituent un événement. L'étiquette de l'événement est attribuée en fonction de l'indice de valeur maximale, de la durée et de caractéristiques complémentaires.
Par exemple si l'indice d'occlusion est prépondérant sur les autres, si le segment est long (≥3 frames) et si l'indice d'acuité est supérieur à un seuil donné, l'étiquette attribuée au segment sera occlusive sourde (Q). Par contre si le segment est bref (2 frames ou moins) et s'il y a accroissement de l'énergie et de l'aigu alors le segment recevra l'étiquette explosion sourde (EX) quand le niveau d'acuité est suffisant sinon explosion voisée (EV).
Un phonème peut se réaliser fréquemment comme une suite d'événements : c'est le cas des occlusives, des vibrantes, des voyelles accentuées et des voyelles nasales longues, etc. Inversement, un événement peut regrouper plusieurs phonèmes: c'est le cas pour des réalisations de clusters tels que [ɔr], [œr], [il], etc. Mais en moyenne le nombre d'événements est presque le double du nombre de phonèmes (à titre indicatif, sur notre corpus le rapport est de 1,8) —voir Fig.1 pour un exemple. On notera que les réalisations observées à la sortie de SAPHO montrent une variabilité par rapport aux modèles idéaux de réalisation. Il conviendra donc de prévoir lors de l'alignement automatique, que des substitutions, des sur-segmentations et des sous-segmentations ont pu se produire.

2.4 - Noyaux syllabiques

Le traitement syllabique consiste ici à associer à chaque événement phonétique la valeur de plausibilité pour qu'il soit le centre d'un noyau syllabique. Le résultat est la création d'un champ syll(i) dont la valeur est 0 si i ne peut être noyau, 1 si i peut être noyau,

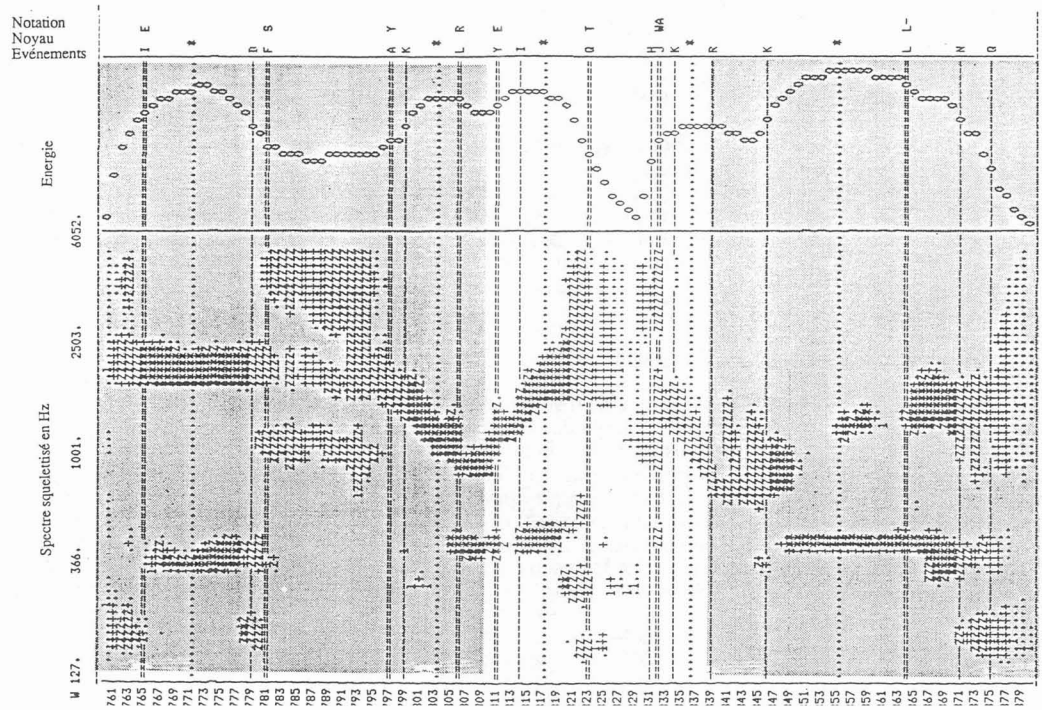

Fig. 1 - Alignement phonémique de la séquence [... syr etwal ...] « sur étoile» de la 3e phrase du corpus EUROM-0.

Le principe d'attribution de la valeur syll(i) est le suivant : lorsqu'on rencontre une suite d'étiquettes de l'ensemble NOY={K,k,I,i,U,u} on les mémorise. On analyse ensuite la séquence mémorisée pour déterminer l'événement prépondérant pour lequel syll(i)=1.
Un segment sera déclaré prépondérant s'il satisfait l'une des propriétés déterminée par expertise, les cas les plus simples étant :

- la liste ne contient qu'un segment (cas le plus fréquent: 60%) qui est alors prépondérant (c'est le cas pour tous les noyaux figurant dans Fig 1),
- la liste comporte trois événements qui sont des noyaux forts: celui du centre est alors considéré comme prépondérant.

Ces critères à eux seuls permettent de localiser correctemment les noyaux vocaliques dans 94% des cas. Ces résultats obtenus par une méthode purement ascendante peuvent être considérés comme bons. On peut envisager de les améliorer par l'introduction de conditions plus raffinées utilisant des comparaisons de durée, d'indices énergétiques ou spectraux. Mais les erreurs restantes posent des problèmes difficiles, voire insolubles et il restera toujours sur un corpus donné des cas où il n'est pas posssible de trancher. Par exemple l'une des erreurs que nous avons relevée se trouvait dans la séquence [swasãt] où le phonème [w] a été déclaré noyau syllabique (syll(i)=1), l'étiquette phonétique k étant isolée. L'examen de la courbe d'énergie et du spectre a montré que le phonème [w] a probablement été réalisé syllabique ([suasãt]), ce que le transcripteur n'avait pas noté.
Pour approcher un taux de syllabation de 100% il faut compléter les traitements de type ascendant par des traitements linguistiques de type descendant, ce qui est le cas lorque nous faisons de l'étiquetage automatique (Cf. 3.2).

3 - L'ALIGNEMENT PHONEMIQUE

Nous allons maintenant examiner comment la structure de la syllabe peut intervenir dans l'étiquetage automatique.

3.1- Le système VERIPHONE

Le système VERIPHONE est décrit dans /11/. Il prend en entrée la chaîne Y d'événements phonétiques (EP) produite par SAPHO et une transcription en unités phonétiques (UP) communiquée par l'expert après écoute du signal. Il produit en sortie la suite des assignations temporelles des UP. Le système VERIPHONE assigne à chaque UP une suite d'EP par application de règles phonétiques fournies par un expert. Voici quelques exemples de règles :

i —> jIl		ar —> yKR	
t —> >Q<	>Q<z	T —> >Q<	>Q<j
+t —> Q<z	Q<z	+T —> Q<	Q<j
t- —> >Q<*			

Un coût est associé à chaque application d'une règle. Celui-ci est évalué en fonction des *opérations* nécessaires pour passer de la suite idéale d'EP figurant en partie droite de la règle à celle, extraite de Y, que VERIPHONE tente de lui faire correspondre.
Les opérations utilisables sont :

- la *fusion* : par exemple j et/ou l peuvent fusionner dans I avec un coût de 15 de telle manière que i —> jIl de coût nul engendre trois autres règles : i —> Il et i —> jI de coût 15, i —> I de coût 30.
- la *substitution* : par exemple la substitution de I par K de coût 15 qui permet de dériver de la règle i —> jIl la règle i —> jKl avec un coût de 15 multiplié par la durée de l'EP K.
- la *sur-segmentation* : si une UP est sur-segmentée le coût de substitution est augmenté de 20.

(N.B. Tous ces coûts sont donnés à titre indicatif: ils font l'objet de règlages dans chaque application particulière).
Une solution optimale d'alignement consiste en une suite de règles, une par UP, de telle manière que: a) chaque EP soit utilisé une et une seule fois, b) que la somme des coûts soit minimale.
La Fig.1 montre un exemple d'alignement où (zone non ombrée) l'on a appliqué la règle t —> >Q< avec *fusion* et *substution* d'où t —> >Q< —> Q< —> Qh.

3.2 - La prise en compte d'informations syllabiques

La contribution des noyaux syllabiques dans l'alignement automatique a été envisagée de deux manières :

- utilisation des noyaux vocaliques détectés pour guider l'alignement,
- traduction de la structure syllabique dans les règles phonétiques.

L'utilisation des noyaux vocaliques pour guider la stratégie d'alignement s'est révélée décevante. En revanche, la prise en compte de la structure syllabique dans la base de règles phonétiques est essentielle. Nous l'avons introduite en distinguant les consonnes d'attaque (+c), intervocaliques (c) ou de coda (c- ou c~). Afin de ne pas multiplier inutilement les règles ceci n'a été fait que lorsque des différences notables dans la structure en événements phonétiques le justifiaient. Nous allons expliquer ceci à partir de l'alignement de la transcription [sw a s an n~ d i s k a t x @ v in s in k an n~ s in- k -] «soixante-dix, quatre-vingt, cinquante-cinq» où l'on utilise les règles suivantes (entre autres):

+n —> bNa		+k —> Q<b	| Q< | QL
n —> dNa	in —> yK~ | yKI~	k —> >Q<	|>Q<b | >QL
n~ —> ~Nc | ~N~	in - —> yKu~ | yKIu~	k - —> >Q<*	

(in- signifie in en syllabe fermée).

Si l'on ne différencie pas suivant la position syllabique nous sommes conduits à poser :
n —> dNa | bNa | ~Nc | ~N~ in —> yK~ | yI~ | yKu~ | y KIu~
k —> Q<b | Q< | QL | >Q< | >Q<b | >QL | >Q<*
puis à faire l'alignement de [sw a s an n d i s k a t x @ v in s in k an n s in k]. Dans ce cas VERIPHONE prend plus de temps pour trouver la solution (20 à 30%). Dans d'autres exemples il peut se produire des erreurs d'alignement que l'on évite en référant les UP à la structure syllabique (c'est-à-dire en prenant la première solution).

4. CONCLUSION

L'utilisation de la syllabe en étiquetage automatique peut contribuer à des améliorations appréciables. Il semble cependant difficile de l'introduire en tant qu'unité de décision dirigeant hiérarchiquement la stratégie d'alignement. Dans notre projet, c'est en attribuant aux unités phonétiques un statut dépendant de leur position par rapport à la structure de la syllabe que les gains les plus importants sont obtenus : augmentation de la qualité des alignements et diminution des temps de calcul d'environ 30%.
Par ailleurs, il ne faut pas sous-estimer les possibilités de contrôles supplémentaires que l'on peut obtenir lorsque l'on peut effectuer un alignement grossier à partir des noyaux syllabiques obtenus de manière ascendante —dans notre cas ces noyaux sont obtenus à partir des événements phonétiques. Cette utilisation de l'information syllabique que nous n'avons pas encore exploitée devrait jouer un rôle important dans l'automatisation complète de l'étiquetage de grands corpus de parole.

5. BIBLIOGRAPHIE

/1/ MELHER, DOUMERGUES, FRAUENFELDER, SEGUI, Journ. of Verb. Learning and Verbal Behav. (1981)
/2/ JHONSON D.H, WEINSTEIN C., IEEE-Trans on Acoustics, Speech, Signal Processing, Vol. ASSP-26, N°5, October 1978, pp. 409-418.
/3/ MERMELSTEIN P., J. Acoust. Soc. Am., Vol. 58, N°4, October (1975).
/4/ LEA W. , IEEE Transactions ASSP-23, (1976) pp. 30-38.
/5 / GAUVAIN J.L., Proceedings IEEE-ICASSP, Tokyo, Vol. 1, (1986) pp. 57-60.
/6/ MORENO A., ARMAS P., MARINO J.B., MASGRAU E., Proceedings European Conference on Speech Communication and Technology 89, Paris, Vol. 2,1989, pp. 75-78.
/7/ MERIALDO B., IBM Journal of R&D, (1988).
/8/ TANAKA A. et al., Speech Communication, Vol 2, Special Issue, 11th ICA, Paris, (1983), pp. 207-210.
/9/ NIST Doc.: Getting Started with the DARPA TIMIT CD-ROM, Gaithersburg, MD:NIST, December 1988.
/10/ SAM—Multi-lingual Speech Input/Output: Assessment, Methodology and Standardisation, Extension Phase, Final Report, 1 April 1988-28 February 1989, pp. 267-279.
/11/ PERENNOU G., J.M. PECATTE, M. DE CALMES, VIGOUROUX N, AFCET, 7e Congrès, Tome 3, Paris, (1989), pp. 1205-1214.