Segmentation automatique de la parole : Pourquoi ? Quels segments ?

Automatic speech segmentation : why and what segments ?

M. ROSSI

URA « Parole et Langage »
Institut de Phonétique
Université de Provence
29 avenue Robert Schuman
13621 Aix-en-Provence
France

Mario ROSSI a obtenu l'Agrégation en 1959, le Doctorat d'Etat en 1974. Il exerce les fonctions de Professeur à l'Université de Provence, de Directeur de l'URA-CNRS 261 « Parole et Langage », de Président du « Centro di Studio per le ricerche di fonetica ». Il est Président du 12e Congrès International des Sciences Phonétiques (Aix-en-Provence, août 1991), Membre du « Permanent Council of Phonetic Sciences » et Vice-Président de l'Association Internationale des Sciences Phonétiques.

RÉSUMÉ

Nous présentons et discutons le modèle SAPHO (segmentation par les connaissances acoustico-phonétiques) mis en œuvre en langage AWK sous UNIX, sur une station de travail Masscomp. Ce système est conçu comme une procédure de segmentation indépendante du locuteur fondée sur une reconnaissance préalable du mode d'articulation phonétique. Dans la plupart des modèles RAP, les connaissances phonétiques sont toujours utilisées, au moins de façon implicite. Elles doivent l'être de façon explicite. Les unités phonémiques ne peuvent pas être directement construites à partir du signal acoustique ; elles ne sont pas encore disponibles à la sortie de SAPHO.

Suivant le modèle de Construction de Niveaux (Level Building), SAPHO fournit un ensemble hiérarchisé de propriétés et de segments acoustiques, de propriétés et de segments phonétiques congruents avec les unités phonétiques et leur structure interne.

La souplesse de ce système est assurée par sa modularité. La fiabilité de SAPHO est corroborée par l'exactitude des résultats.

MOTS CLÉS

acoustique, connaissances, macroclasses, mode d'articulation, phonémique, phonétique, reconnaissance, segments, segmentation automatique.

SUMMARY

I present and discuss the SAPHO (Segmentation by Acoustico-Phonetic knowledge) model implemented in Awk language under the Unix system on a MASSCOMP computer. The system is devised as a speaker independent ASS (automatic speech segmentation), by a previous recognition of the phonetic articulation manner. In all the ASR systems the phonetic knowledge is at least implicitely used. It has to be explicitely referred to. The phonemic units cannot be directly built from the acoustic signal and are not available at the output of SAPHO. According to the Level Building procedure SAPHO supplies a hierarchized set of acoustic properties and segments, and phonetic properties and segments which fit the phonetic parsing of the acoustic wave. The amenability of this system is entailed by its modularity which allows a possible further architecture as distributed tasks.

The processors are concieved either as data driven with numeric computation or as expectation driven activities with symbolic computation. The recursivity in the acoustic and the phonetic supervisors at each step of the parsing ensures the likelihood of the decisions. The suitability and the reliability of SAPHO are corroborated by the accuracy of the results.

KEYWORDS

acoustics, knowledge, macroclasses, manner of articulation, phonetics, phonemics, recognition, segments, automatic segmentation.

Introduction

De nombreux chercheurs ont utilisé et utilisent encore les procédures de segmentation dans la reconnaissance automatique de la parole (RAP). D'autres suppriment cette étape. Deux questions se posent de prime abord.

1) Étant donné la rupture de biunivocité entre les niveaux linguistique et acoustique, est-ce que la segmentation automatique est justifiée ? Si oui, les procédures de segmentation fournissent-elles une information utile ou au contraire risquent-elles d'égarer la reconnaissance sur de fausses pistes ?

2) Si à cette première question nous répondons que la segmentation automatique est justifiée et fiable, quels types d'unités celle-ci permet-elle d'isoler (segments, phones, phonèmes ?), en d'autres termes quel niveau la segmentation permet-elle d'atteindre ?

Des réponses correctes à ces questions permettraient de nous éclairer sur l'organisation d'un module de décodage acoustico-phonétique (DAP). Nous avons élaboré un système de segmentation du continuum acoustique sur la base de connaissances phonétiques (SAPHO) ; SAPHO conduit à la reconnaissance de macroclasses phonétiques. La présentation et la discussion de ce système nous permettra d'apporter des éléments de réponse aux questions posées ci-dessus.

1. Fondements théoriques

Beaucoup de modèles RAP ont été ou sont encore fondés sur les résultats d'un module de segmentation automatique (SA) qui est supposé construire directement des unités phonologiques, les phonèmes, à partir des paramètres acoustiques. Ces procédures de segmentation automatique conduisent à des erreurs fatales et ont été justement critiquées au cours des dix dernières années [12, 14] : ces procédures SA sont en contradiction avec la théorie linguistique et ce que nous savons du traitement du langage par l'humain. Afin d'aborder correctement les problèmes concernant la segmentation de la parole en général, il est nécessaire d'avoir une idée claire des relations qui s'établissent entre les différents niveaux du traitement de la parole.

1.1. On sait qu'il n'existe pas de correspondance biunivoque entre la représentation phonémique abstraite et les événements acoustiques et articulatoires concrets. En particulier, les phonèmes sont des unités dont le caractère discret n'a aucune existence dans l'onde acoustique. Le continuum acoustique est phonologiquement opaque : il ne nous fournit aucune indication, par exemple, pour distinguer les énoncés anglais *grey tape* et *great ape* [17].

Mais la question se pose de savoir si la rupture de biunivocité s'étend également aux relations qui lient les plans phonétique et acoustique.

Le plan phonétique a longtemps été confondu avec les niveaux concrets, acoustiques et articulatoires ; en réalité il doit être clairement distingué de ces deux derniers : il représente en effet un niveau de représentation intermédiaire entre les événements concrets et la représentation phonologique.

Mais il n'existe pas davantage de correspondance biunivoque entre les représentations phonologique et phonétique. En français, par exemple, à partir de la représentation phonétique [o], il n'est pas possible de recouvrer l'unité phonologique qu'elle réalise, /ə/ ou /ɔ/, si nous n'avons pas recours à la connaissance linguistique concernant la modification du schwa dans le contexte de la consonne /*R*/. De la même manière, la représentation phonétique du français [S: yi] ne nous permet pas d'accéder aux phonèmes /ʒ/ et /s/ de la suite [ʒ ə— sy i] ; à moins de connaître la règle de syncrétisme de ces deux consonnes, dans un certain contexte, en une consonne palato-alvéolaire longue qui appartient au niveau de la représentation phonétique.

Ces exemples montrent que l'accès aux unités phonologiques à partir du niveau phonétique exige des connaissances linguistiques, en particulier les règles de réalisation et de représentation intermédiaires. A fortiori, il tombe sous le sens que l'accès aux unités phonologiques à partir des événements acoustiques concrets est impossible. Les événements acoustiques concrets en effet sont en relation directe avec les unités phonétiques, non avec les unités phonologiques. Dans l'exemple ci-dessus, ce n'est pas la séquence /ʒ + s/, mais la représentation phonétique [S :] qui est transférée dans le continuum acoustique concret : ce qui ne signifie pas, bien entendu, qu'il existe un rapport nécessaire de biunivocité entre ces deux derniers niveaux.

Les unités du niveau phonétique sont également caractérisées par leur caractère discret. Mais parce que les règles d'organisation, en d'autres termes parce que la logique des plans acoustique et phonétique est différente, le caractère discret des unités phonétiques ne peut pas être conservé dans le continuum acoustique.

Pourtant des discontinuités évidentes apparaissent dans le signal acoustique ; mais quelle est leur valeur phonétique ? Afin d'identifier certaines de ces discontinuités comme le passage d'une occlusive à une voyelle, nous devons connaître et la structure phonétique des occlusives et des voyelles et les règles de transformation du niveau phonétique aux niveaux articulatoire et acoustique.

Par conséquent, l'accès aux unités abstraites de haut niveau est loin d'être direct et exige une connaissance préalable complexe du codage de la parole. Cette exigence est admise par de nombreux chercheurs tant phonéticiens qu'ingénieurs [7, 14, 18, 24].

« It is our belief that acoustic-phonetic knowledge representation is the major roadblock in the design of advanced speech recognition systems that are meant to approach human performance », écrit Zue en 1982.

Cette affirmation de Zue semble être une évidence. Pourtant un certain nombre de chercheurs, et non des moindres, devant l'efficacité des modèles de Markov et des réseaux de neurones en RAP laissent entendre que les systèmes à base de connaissances sont maintenant dépassés et doivent être abandonnés parce qu'ils n'auraient prouvé que leur inefficacité.

Toutefois certains chercheurs qui développent des réseaux de neurones pour la reconnaissance automatique soulignent la nécessité d'une segmentation préalable précise pour pouvoir aligner correctement les entrées du signal acoustique et les réseaux de neurones [23].

D'autres [9] ont montré l'avantage qu'on pouvait tirer de connaissances phonétiques explicites dans le fonctionnement des reconnaisseurs phonétiques HMM.

Il va sans dire que nous ne pourrons pas résoudre les problèmes majeurs de RAP, notamment pour la parole continue, si nous n'avons pas recours à notre connaissance de la structuration du langage et de l'encodage phonologi-

que aux niveaux phonétique et acoustique. Tous les systèmes RAP, de façon explicite ou implicite, utilisent cette connaissance qui gagnerait à être explicitement définie.

1.2. A ce point de notre discussion, un nouveau problème surgit. Une lecture attentive de F. de Saussure, le grand linguiste génevois, montre que le concept de Langue est un composé de matière (substance) et de forme, que la matière (les événements acoustiques) n'est pas amorphe, contrairement à ce qui a été avancé par certains linguistes, mais qu'elle est en dernière analyse organisée par la forme (c'est-à-dire le plan phonémique). Toutes les recherches sur l'acoustique de la parole, depuis des années, ont mis en évidence l'organisation complexe de ce qu'on appelle le continuum acoustique.

Alors pourquoi les événements acoustiques ne donnent-ils pas accès directement aux unités phonologiques ? Parce que la logique de l'organisation de ces deux plans est tout à fait différente et sans commune mesure. Pour poser le problème en termes clairs, nous pouvons admettre, comme en Physique, que nous avons une théorie et des faits empiriques liés par une interface qui a pour objet de saisir et d'analyser les faits à la lumière de la théorie et d'évaluer l'adéquation de la théorie à l'explication des faits. Mais contrairement à la Physique, l'interface, dans le traitement de la parole, a une puissance qui est renforcée par les modèles de contrôle que sont les modèles du locuteur et de l'auditeur (*fig.* 1).

Cette conception de l'organisation des domaines de la Phonétique et de la Phonologie devrait prouver son efficacité dans le domaine des technologies vocales. Elle évite le vieux divorce entre Phonétique et Phonologie et permet une distinction claire entre les niveaux de représentation et les faits empiriques. Afin d'éviter toute querelle de clocher, j'appellerais volontiers la science de ce domaine, la science PH qui active, dans l'interface, les méthodes de la Phonétique, de la Phonologie et de la Psychologie.

	Forme	Structure	Théorie	Auditeur	
Langue	|	|	INTERFACE	Modèles	Technologies de la parole
	Substance	Matière organisée	Faits empiriques	Locuteur	

Fig. 1.

Ce modèle nous indique clairement que toute tentative pour atteindre les unités de haut niveau à partir des seuls faits empiriques est vouée à l'échec. Les faits empiriques, c'est-à-dire les événements acoustiques, doivent être interprétés par l'interface.

1.3. Si nous considérons la complexité de l'organisation des événements acoustiques et l'effacement du caractère discret des unités phonémiques par le processus de coarticulation, surgit alors une nouvelle question : est-ce que l'information contenue dans les faits empiriques contient des indices qui permettraient d'identifier des unités abstraites de type phonologique ? En d'autres termes, peut-on avoir accès à des unités de haut niveau par des procédures de segmentation. Et quelle que soit la réponse à cette question, une segmentation a priori est-elle souhaitable, sinon comment procéder ?

Les travaux sur la coarticulation montrent que ce phénomène de coproduction met en jeu essentiellement, et en premier lieu, les paramètres et les indices qui codent le lieu d'articulation. Les paramètres impliqués dans le codage du mode d'articulation échappent en partie à l'effet de chevauchement de la coarticulation. Le score d'intelligibilité élevé des traits de mode d'articulation dans les tests par paires minimales peut être considéré comme une preuve de ce fait. En conséquence, la résistance relative des paramètres de mode d'articulation nous autorise à segmenter le soi-disant continuum acoustique. Nous devons cependant être conscients du fait que les segments obtenus sur la base des paramètres acoustiques ont un domaine sub-phonémique ou supraphonémique ; et que le traitement et l'interprétation corrects de ces paramètres et de ces segments doivent être guidés par la connaissance présente dans l'interface. On ne peut avoir accès aux unités phonémiques du dernier niveau si on ne s'appuie pas sur une stratégie descendante explicite qui fasse appel aux contraintes phonémiques et lexicales.

Les discussions sur la hiérarchie et la nature des segments obtenus sur une base purement acoustique présentent un intérêt mineur. En effet, les résultats de la segmentation dans ce cas dépendent des choix effectués dans le système RAP et par conséquent des seuils qu'on s'est fixés dans un certain but. Les segments ainsi obtenus ne sont pas forcément congruents aux unités phonétiques et en un certain sens sont arbitraires.

Il semble donc qu'une segmentation a priori directement fondée sur des paramètres tels que l'énergie et la dérivée du spectre, à partir d'un seuil prédéfini, n'est pas fiable et risque de dévoyer les processus de reconnaissance vers de fausses pistes. Les systèmes RAP, dans ce cas, dépendent exagérément de frontières strictes, arbitrairement fixées entre les segments.

D'autres méthodes beaucoup plus élaborées, telles que la décomposition temporelle, ne donnent pas toujours les résultats escomptés. Bien que le plus souvent elles utilisent des coefficients articulatoirement fondés, ces techniques de décomposition temporelle en principe ne font appel à aucune connaissance phonétique explicite. Aussi certains auteurs [22] proposent-ils d'améliorer la performance de ces méthodes par l'introduction, en particulier, de connaissances phonétiques.

2. Description du système

En accord avec ce que nous avons dit plus haut concernant les rapports entre les niveaux phonologique, phonétique et acoustique nous voudrions ici présenter un modèle de reconnaissance automatique de macroclasses, plus précisément d'identification de segments acoustiques construits et

interprétés par un système à base de connaissances. Ces segments sont congruents à la structure des unités phonétiques ; ils sont obtenus par une procédure de segmentation indirecte fondée sur des propriétés des échantillons de parole.

Toutes les connaissances auxquelles il est fait appel sont orientées vers la reconnaissance exclusive des classes phonétiques de mode d'articulation (occlusif, constrictif, etc...). La phase de reconnaissance des lieux d'articulation, actuellement en cours de développement, n'est pas exposée ici. Au cours de cette seconde phase, le résultat de la segmentation de SAPHO sera mis en correspondance avec l'organisation temporelle des indices et des traits de lieu d'articulation ainsi qu'avec les mouvements de certains articulateurs. La position de ce modèle par rapport aux théories articulatoires pourra alors être précisée de façon opportune.

Ce modèle SAPHO est conçu comme une première étape d'un système RAP indépendant du locuteur. Il est mis en œuvre à l'aide d'un programme modulaire écrit en langage AWK (1) sous UNIX, sur une station de travail Masscomp.

Le système SAPHO, dont la stratégie est spécifiée dans le paragraphe 4, vise dans une première étape à identifier les classes fondamentales VOYELLE et CONSONNE à partir de propriétés de base robustes qui portent les étiquettes CSB (consonne, silence ou bruit), FRIC (fricatif) et VOC (Vocalique). Ces propriétés sont ensuite interprétées par des connaissances acoustiques, phonétiques et contextuelles pour créer des segments subphonétiques ou phonétiques tels que SIL (silence), EXPL (explosion), LIQ (organisation de consonne liquide L ou N), VOY (partie croissante de voyelle), VOX (partie décroissante de voyelle) etc...

En dernier ressort, des règles phonologiques créent les classes phonétiques VOYELLES, CONSONNES VOCALIQUES, CONSTRICTIVES, OCCLUSIVES etc...

2.1. PARAMÈTRES ET INDICES

Nous utilisons, dans une première étape, un corpus de 20 mots prononcés par 3 locuteurs (deux hommes, et une femme) pour tester les modules et le noyau.

Les fichiers de données contiennent les spectres obtenus toutes les 10 ms par un vocodeur à 15 canaux (note 1). La contribution de chaque bande à l'organisation acoustique est représentée par le pourcentage de l'énergie dans chaque canal (12).

2.1.1. Les primitives calculées sur chaque échantillon sont :

— la densité de passages par zéro : PZ
— l'énergie totale (somme des énergies dans chaque canal) : EN
— la dérivée première de EN (différence algébrique entre deux échantillons successifs) : DE
— la dérivée première du spectre calculée sur les contributions (somme des différences absolues canal à canal entre deux échantillons successifs) : DK
— l'énergie dans les basses fréquences (somme de l'énergie dans les canaux 1 à 4) : BF
— l'énergie dans les hautes fréquences (somme de l'énergie dans les canaux 9 à 14) : HF

Les paramètres EN%, BF% et HF% représentent respectivement les primitives EN, BF et HF normalisées.

2.1.2. Les indices sont testés sur la distribution significative de l'énergie dans le spectre. Les programmes d'indices prennent la forme de règles d'action ; par exemple, A2 qui teste la présence de l'indice AIGU2 :

```
{if  (K7 + K8 >  K5 + K6)
   A2 = « + » ; print
 else
   A2 = « – » ; print}
```

La pertinence de ces tests est fondée sur les connaissances phonétiques.

La plupart des indices utilisés dans TRAITS2 ont été définis par M. Rossi [20, 21], modifiés et mis en œuvre sous le système Expert IROISE-SERAC du CNET par A. Bonneau [3].

Les traits d'ouverture, d'acuité et de bémolisation sont évalués selon deux procédures. Les tests sur les indices de ces traits constituent des modules indépendants interprétés par deux processeurs : TRAITS2 et TRAITS3.

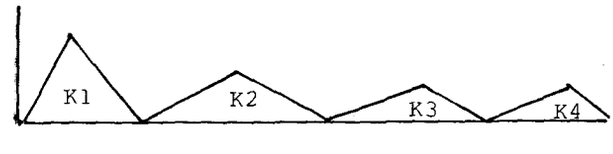
Le premier est construit sur les réponses binaires aux tests, le second sur les réponses floues. Le contenu des indices est également différent dans chacun des processeurs.

Dans le processeur binaire TRAITS2, on calcule respectivement 1, 4 et 10 indices pour les traits de bémolisation, d'ouverture et d'acuité. Pour chaque trait, les indices sont hiérarchisés et un coefficient attribué à chaque indice en fonction de sa place dans la hiérarchie (Tableau 1). Sur chaque ligne, pour un échantillon, la combinatoire des indices fournit le profil et la valeur numérique du trait présumé. Afin d'éviter une combinatoire trop importante pour le trait d'acuité (10 indices !), les indices sont répartis en 3 catégories. Le mode de calcul que nous venons de présenter est appliqué à chacune de ces catégories.

Quelques exemples permettent de comprendre l'intérêt de cette procédure.

La combinatoire des indices pour chaque trait (ouverture, acuité) détermine une valeur numérique dépendante du nombre d'indices : de 1 à 16 (2^4) et de 1 à 8 (2^3). Le caractère binaire des indices évite le recours à des seuils particulièrement indésirables dans une reconnaissance multilocuteurs et montre son efficacité dans l'identification des macroclasses. La combinatoire des indices présente l'intérêt de saisir des structures : ainsi l'indice numérique qui accompagne chaque combinaison n'est pas une valeur scalaire mais l'image d'une organisation complexe qu'il est difficile d'appréhender directement. Ainsi la valeur 16 pour le trait d'ouverture (Tableau 1, colonne 1) représente

la combinaison – – + + et le spectre



La valeur 10, pour le même trait, représente la combinaison − − + − et le spectre

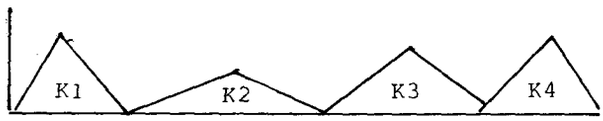


TABLEAU 1

Sortie du processeur TRAITS2 *sur le mot* beurre, *locuteur femme V.A. Combinaisons binaires des indices acoustico-phonétiques. Pour chaque échantillon, la combinaison est suivie d'une valeur numérique qui fournit une évaluation du trait concerné. Trait d'ouverture (4 indices), trait d'acuité (10 indices : 3 + 3 + 4), trait de bémolisation.*

TRAITS BINAIRES sur beurre.va

NIECH	OUVERTURE		ACVITÉ1		ACVITÉ2		ACVITÉ3		BÉMOLISATION
1	−−−−	3	−−−	8	−−−	1	−−−−	16	−BEM. 1
2	+−−−	1	−−−	8	−−−	1	−−−−	16	−BEM. 1
3	+−−−	1	−−−	8	−+−	3	−+−−	5	+BEM. 1
4	−−++	16	−−−	8	−+−	3	++−−	4	+BEM. 1
5	−−++	16	−−−	8	−+−	3	−+−−	5	+BEM. 1
6	−−++	16	−−−	8	−+−	3	−+−−	5	+BEM. 1
7	−−++	16	−−−	8	−+−	3	+++−	1	+BEM. 1
8	−−++	16	−−−	8	++−	5	++−−	4	+BEM. 1
9	−−++	16	−−+	7	+−−	6	−−−−	16	+BEM. 1
10	−−++	16	−−−	8	−−+	2	−−−−	16	−BEM. 1
11	−−++	16	++−	2	−−+	2	−−−−	16	−BEM. 1
12	−−++	16	−−−	8	+−+	7	−−−+	15	+BEM. 1
13	−−+−	10	−−−	8	++−	5	−−−+	15	+BEM. 1
14	+−−−	1	−−+	7	+++	8	−−−−	16	+BEM. 1
15	+−−−	1	−−−	8	+−+	7	−−−+	15	+BEM. 1
16	+−−−	1	−−+	7	+−+	7	−−−+	15	+BEM. 1
17	+−−−	1	−−+	7	+−+	7	−−−+	15	+BEM. 1
18	+−−−	1	+−+	3	+−+	7	−−−+	15	+BEM. 1
19	+−−−	1	+−+	3	+++	8	−−−+	15	+BEM. 1
20	+−−−	1	+++	1	+++	8	++−+	3	+BEM. 1
21	+−−−	1	−−−	8	+++	8	++−+	3	+BEM. 1
22	+−−−	1	−−+	7	+++	8	−−−−	16	+BEM. 1
23	+−−−	1	−−+	7	+++	8	−−+−	9	+BEM. 1
24	+−−−	1	−−−	8	+++	8	−−+−	9	+BEM. 1
25	+−−−	1	−−+	7	++−	5	−−+−	9	+BEM. 1
26	−+−−	8	−−+	7	+++	8	−−−+	15	+BEM. 1
27	−+−−	8	−−+	7	++−	5	−−−+	15	+BEM. 1
28	−+−−	8	−−+	7	+−−	6	−−−+	15	+BEM. 1
29	−+−−	8	−−+	7	++−	5	−−−+	15	+BEM. 1
30	−+−−	8	−−−	8	−+−	3	−−−+	15	+BEM. 1
31	+−−−	1	−−−	8	−+−	3	−−−+	15	+BEM. 1
32	+−−−	1	−−+	7	++−	5	−−−+	15	+BEM. 1
33	+−−−	1	−−−	8	++−	5	−−−−	16	+BEM. 1
34	+−−−	1	−−−	8	−+−	3	−−−−	16	+BEM. 1
35	+−−−	1	−−−	8	++−	5	−−−−	16	−BEM. 1
36	+−−−	1	−−−	8	+−−	6	−−−+	15	−BEM. 1
37	+−−−	1	−−−	8	+−+	7	−−++	10	+BEM. 1
38	+−−−	1	−−−	8	+−−	6	−−−+	15	−BEM. 1
39	+−−−	1	−−−	8	+−−	6	−−−+	15	−BEM. 1
40	+−−−	1	−−−	8	+−−	6	+−++	7	−BEM. 1
41	+−−−	1	−−−	8	+−−	6	−−−+	15	−BEM. 1
42	+−−−	1	−−−	8	+−+	7	−−−+	15	+BEM. 1
43	−−+−	10	−−−	8	−+−	3	−−++	10	+BEM. 1
44	−−+−	10	−−−	8	++−	5	+−+−	6	−BEM. 1
45	−−+−	10	−−−	8	−+−	3	+−++	7	+BEM. 1
46	−−+−	10	−−−	8	−+−	3	−−−+	15	+BEM. 1
47	+−−−	1	−−−	8	++−	5	−−−+	15	+BEM. 1
48	−−+−	10	−−−	8	−+−	3	−−−+	15	+BEM. 1
49	−−+−	10	−−−	8	−+−	3	−−−+	15	+BEM. 1
50	−−+−	10	−−−	8	++−	5	−−−+	15	+BEM. 1

Cette structuration de formes complexes autorise une comparaison relativement simple entre traits, par exemple entre le trait d'ouverture (Tableau 1, colonne 1) et le premier trait d'acuité (Tableau 1, colonne 2). Ainsi les valeurs 1 (ouverture) et 1 (acuité) qui représentent les combinaisons respectives

\+ − − − (tests sur les canaux 1 à 3)
et + + + (tests sur les canaux 4 à 6)

renvoient à un spectre spécifique de voyelle. On devine

TABLEAU 2

Sortie du processeur TRAITS3 *sur le mot* beurre, *locuteur femme V.A. Analyse floue des indices acoustico-phonétiques pour le trait d'ouverture (colonnes 2, 3, 4), de gravité (colonnes 5, 6, 7) et d'acuité (colonnes 8, 9, 10).*

TRAITS FLOUS sur beurre.va

1	F0	—	—	—	G5	—	—	A10	—
2	—	—	07	G2	—	—	—	A10	—
3	—	—	06	—	—	G5	—	A9	—
4	F1	—	—	—	—	G5	—	A9	—
5	F1	—	—	—	—	G5	—	A10	—
6	F1	—	—	—	—	G5	—	A9	—
7	F1	—	—	—	—	G5	—	A9	—
8	F1	—	—	—	—	G5	A9	—	—
9	F1	—	—	—	G5	—	A9	—	—
10	F1	—	—	—	G5	—	—	—	A9
11	F1	—	—	—	G5	—	—	—	A9
12	F2	—	—	G2	—	—	A9	—	—
13	F4	—	—	G2	—	—	A9	—	—
14	—	—	06	G2	—	—	A9	—	—
15	—	—	07	G2	—	—	A9	—	—
16	—	—	07	G2	—	—	A8	—	—
17	—	—	06	G2	—	—	A8	—	—
18	—	05	—	G2	—	—	A8	—	—
19	—	05	—	G2	—	—	A7	—	—
20	—	05	—	—	—	G8	A7	—	—
21	—	05	—	—	—	G6	A9	—	—
22	—	—	07	G3	—	—	A8	—	—
23	—	—	07	—	G4	—	A8	—	—
24	—	—	07	—	G4	—	A9	—	—
25	—	—	06	—	G4	—	A8	—	—
26	—	05	—	G2	—	—	A8	—	—
27	—	05	—	G2	—	—	A8	—	—
28	—	05	—	G2	—	—	A8	—	—
29	—	05	—	G2	—	—	A8	—	—
30	—	05	—	G2	—	—	—	A9	—
31	—	—	06	G2	—	—	—	A9	—
32	—	—	06	G2	—	—	A9	—	—
33	—	—	07	G2	—	—	A9	—	—
34	—	—	07	G2	—	—	—	A9	—
35	—	—	07	G3	—	—	A9	—	—
36	—	—	07	G2	—	—	A9	—	—
37	—	—	06	—	G4	—	A9	—	—
38	—	05	—	G2	—	—	—	—	—
39	—	—	07	G3	—	—	—	—	A13
40	—	05	—	—	G4	—	A9	—	—
41	—	—	07	G2	—	—	A9	—	—
42	—	—	06	G3	—	—	A9	—	—
43	F3	—	—	—	G4	—	—	A9	—
44	F3	—	—	—	G4	—	A9	—	—
45	F4	—	—	—	G4	—	—	A9	—
46	F4	—	—	G2	—	—	—	A9	—
47	—	—	07	G2	—	—	A9	—	—
48	F3	—	—	G2	—	—	—	A9	—
49	F2	—	—	G2	—	—	—	A9	—
50	F2	—	—	—	G4	—	A9	—	—

tout l'intérêt de cette procédure pour le traitement de spectres complexes, la reconnaissance des voyelles, l'identification des macroclasses et la segmentation.

Dans le processeur flou TRAITS3, les tests définissent 9 indices : 3 pour le trait d'ouverture, 6 pour le trait d'acuité. Chaque test recherche le maximum d'énergie dans certaines bandes et calcule la valeur de l'énergie dominante (ED).

Selon la valeur de ED et suivant certaines conditions, chaque trait est représenté par un symbole sur 3 échelles à 4 degrés (12 degrés au total). Ainsi pour le trait d'ouverture (Tableau 2) :

Echelle 1 = *fermé,* degrés F0, F1, F2, F3
Echelle 2 = *moyen,* degrés F4 etc...
Echelle 3 = *ouvert,* degrés 06 etc...

Alors que dans TRAIT2 les indices constituent des tests spécifiques sur des bandes qui ne sont pas forcément contiguës, dans TRAIT3 on recherche les maxima d'énergie dans trois bandes du spectre : F/O (fermeture/ouverture), G (gravité) et A (acuité). Chaque maximum est pondéré par son degré d'émergence. On obtient ainsi un spectre catégorisé sur des valeurs floues. Cette représentation est un complément indispensable des résultats de TRAITS2 pour la reconnaissance fine des voyelles par exemple, mais aussi pour l'identification des classes de consonnes : la catégorie F0 dans la première bande, par exemple, indique sans ambiguité la présence d'une consonne occlusive ou constrictive non voisée.

2.2. TRAITEMENT DES PARAMÈTRES ET DES INDICES

Les paramètres sont calculés par des programmes indépendants appelés et organisés par le processeur PARAM. Après le calcul des paramètres, la reconnaissance des macroclasses fait appel à 13 modules indépendants.

Les 7 premiers modules traitent l'information acoustique sur des critères tirés des connaissances phonétiques ; ils sont hiérarchisés et appelés récursivement par un superviseur acoustico-phonétique appelé MODE (Tableau 3). Les autres modules, au nombre de 6, sont appelés après MODE par le superviseur phonétique SUPERMODE (tableau 4).

Les derniers modules dans SUPERMODE sont essentiellement des processeurs à base de connaissances phonétiques qui sont activés par des requêtes [7]. A chaque étape, les résultats intermédiaires sont préservés ; ainsi l'histoire des structurations fournies par les processeurs est disponible à chaque instant dans la mémoire immédiate pour une interprétation ultérieure éventuelle ou une réévaluation (Tableau 5). Cette organisation est apparentée, malgré la différence de conception des systèmes, aux représentations multi-niveaux [8].

L'organisation des modules de SAPHO peut être représentée par l'arborescence suivante, où : VOC = classe vocalique, MAX/MIN = parties croissantes et décroissantes de VOC, les parenthèses insèrent le nom des modules :

PARAM	EN, EN% ..... DK PZ
MODE	Calcul Propriétés de base (RZ)
	Calcul des MAX/MIN de Voc (KL3)
	Test de durée sur MIN/VOC (TRO)
	Interprétation acoustico-phonétique des MAX/MIN (KL4 à KL6)
SUPERMODE	Identification des segments phonétiques (KL8 à KL11)
	SORTIE

2.2.1. Des traits robustes et des îlots de confiance existent bien dans le signal de parole [14]. Aussi la première étape met-elle en œuvre un module, RZ, qui recherche les classes acoustiques CSB, FRIC et VOC déterminées par la robustesse de certains indices. CSB est identifiée de façon

TABLEAU 3

Programme MODE qui appelle PARAM (calcul des paramètres) et les modules acoustico-phonétiques

```
pas = $HOME/bin
for i
do
#echo " "
#echo "FICHIER :$i"
#echo " "
#echo "                BF    EN    DE    EN%    BF    HF    HF%    ZER    DK    K1    PZ
RZ    K3    TRO    K3    K4    K5    K6    TRO    K3    K4    K5    K6"
#echo " "
param $entree$i | $pas/erz | $pas/enet | $pas/ek13> bide4
<bide4 $pas/etrop | paste bide4 — | $pas/epource | $pas/ek13> bide5
<bide5 $pas/etrop | paste bide5 — | $pas/einter | $pas/epource | $pas/ek13 \
       | $pas/eras1 | $pas/einter | $pas/ek14 | $pas/ek15 | $pas/ek16> bide6
<bide6 $pas/etrope | paste bide6 — | $pas/eintezer | $pas/epource | $pas/ek13 \
       | $pas/eras1 | $pas/eintezer | $pas/ek14 | $pas/ek15 | $pas/ek16 \
       | $pas/format3 | cat -n
done
```

TABLEAU 4

Superviseur SUPERMODE qui appelle le programme MODE et les processeurs phonétiques pour la dérivation des segments phonétiques et des macroclasses.

```
# prog pour reconnaissance du mode et du lieu des consonnes
for  i
do
echo " "
echo " "
echo "FICHIER $i"
echo " "
protete
echo " "
mode $entree$i > bide11
traits2 $entree$i > bide12
traits3 $entree$i > bide13
paste bide12 bide13 bide11 | ek18 \
| ek19 | ekl10 | ekl11 | format
#rm bide11 bide12 bide13
done
```

à être congruente avec la classe des consonnes occlusives. La présence de zéros significatifs sur le spectre peut être considérée comme un indice acoustique fiable de ce trait consonantique. Le nombre de zéros est représenté par le nombre de canaux dont la contribution spectrale est nulle. Le produit de EN par le nombre pondéré de zéros fournit une valeur négative qui permet d'identifier les échantillons CSB. L'étiquette CSB sera interprétée plus tard soit comme une consonne occlusive, soit comme un silence. L'étiquette FRIC est dérivée des valeurs de PZ si certaines conditions sont vérifiées lors de la comparaison de EN avec PZ et DK. FRIC indique la présence d'un bruit de forte énergie qui pourra être interprété plus tard soit comme une constrictive (fricative) soit comme un bruit d'explosion. L'étiquette VOC définit tous les échantillons non étiquetés CSB ou FRIC.

Le choix d'une étiquette n'est jamais naïf. Ainsi VOC rappelle le trait Vocalique. Nous savons en effet que d'une certaine manière les consonnes voisées possèdent quelques unes des propriétés des voyelles, elles sont vocalisées.

Il convient de souligner qu'à cette étape la segmentation qui résulte de l'étiquetage n'est pas fondée sur le choix arbitraire préalable d'un seuil de la fonction d'instabilité par exemple, mais sur des paramètres acoustiques spécifiques interprétés par les connaissances phonétiques.

2.2.2. A l'étape suivante, le module KL3 calcule les parties montantes et descendantes de EN. Les parties montantes sont représentées par une suite de valeurs numériques croissantes établies à partir de EN, les parties descendantes par les étiquettes **min1** et **min2. min1** et **min2** représentent deux degrés de décroissance distingués sur la base d'une comparaison entre EN et DE. Les séquences croissantes de EN contiennent les maxima significatifs d'énergie : ils indiquent, avec les portions **min1,** la présence d'une voyelle. Mais les étiquettes **min2** devront être réinterprétées soit comme une partie décroissante de la voyelle soit comme une consonne.

2.2.3. Le module KL3 traite récursivement les séquences **min2** si elles dépassent la longueur moyenne des voyelles dans des mots isolés. A ce stade, KL3 calcule les parties croissantes et décroissantes successivement à partir d'une normalisation locale de EN et de BF, selon la même procédure que précédemment. La normalisation sur le mot entier introduit un lissage important ; la normalisation locale permet de faire émerger les maxima et les minima omis lors du premier passage. Le choix de BF après EN se justifie lors d'un troisième passage par la relation qui est censée exister entre les voyelles et les consonnes qui restent théoriquement à identifier : on peut faire l'hypothèse en effet qu'à ce stade les consommes qui restent à identifier sont les consonnes vocaliques (latérales, nasales ou glissantes). La base de connaissances phonétiques nous dit que dans les portions du signal de faible énergie, les voyelles et les consonnes vocaliques se distinguent essentiellement par des changements dans une bande de basse fréquence. A la sortie de KL3, les longues séquences **min2** sont chaque fois segmentées en parties croissantes et décroissantes si les tests de KL3 sur EN ou BF l'autorisent.

2.2.4. Après le troisième passage de KL3 le superviseur MODE appelle successivement les modules KL4, KL5 et KL6. Ces modules fournissent une interprétation phonétique des séquences acoustiques croissantes (valeurs numériques croissantes) et décroissantes (**min1** et **min2**) KL4 et KL5 sont des processeurs acoustico-phonétiques (interprétation phonétique de séquences acoustiques) dont la décision dépend de règles dépendantes du contexte et qui de ce fait opèrent de droite à gauche en remontant le signal. Les séquences numériques croissantes et les étiquettes **min1** sont reconnues comme des voyelles et étiquetées respectivement VOY (tension de la voyelle) et VOX (détente de la voyelle). Les séquences **min2** sont interprétées soit comme CS (consonnes), soit comme LIQ (liquides), soit comme VOX (détente de voyelle). Les séquences [LIQ VOX LIQ] pourront être ultérieurement identifiées soit comme des latérales, soit comme des nasales. L'étiquette CS fait référence à des consonnes plus faibles que celles qui sont représentées par CSB ; elle pourra être interprétée soit comme obstruente voisée, CVS (occlusive voisée, constrictive voisée, certaines variantes de R), soit comme consonne vocalique, CV.

2.2.5. A l'étape suivante, on recherche les lieux d'omission possible des consonnes. Le module KL6 traite les séquences VOX sur des critères sévères qui font intervenir une comparaison entre EN, DE et DK. Ensuite, on recherche les longues séquences VOX dérivées de **min2** sur lesquelles les critères précédents fondés sur EN% ou BF% n'ont pu identifier de consonnes. Cette fois-ci KL3 interprète les séquences VOX sur la base de RZ, c'est-à-dire de l'énergie totale corrigée par les zéros spectraux. En effet EN% et BF% dont la normalisation lisse fortement l'évolution de l'énergie dans le temps ne donnent plus de résultats. KL3 identifie les parties croissantes et décroissantes.

Les séquences décroissantes étiquetées min2 sont traitées récursivement par KL4, KL5 et KL6 dont nous avons vu plus haut le rôle ; les consonnes identifiées à la sortie de ces modules sont interprétées comme des consonnes

TABLEAU

Résultats de SAPHO sur le mot beurre, *locuteur femme V.A.*
Pour simplifier l'écriture, les noms des modules Klx sont représentés par Kx.

FICHIER beurre.va

N	EN%	EN	DE	BF	BF%	HF	HF%	ZER	DK	K1	PZ	RZ	K3	TRO	K3	TRO	K3	K4	K5	K6	TRO	K3	K4	K5	K6	K8	MOD
1	0	0	0	0	0	0	0	-60	0	0	0.02	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	SIL
2	0	0	0	0	0	0	0	-55	97	18	0.03	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	SIL
3	1	7	1	2	0	1	0	-3	113	7	0.05	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	SIL
4	0	6	-1	4	0	0	0	-34	125	71	0.05	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	OC+VX
5	2	14	2	12	2	0	0	-36	20	77	0.05	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	OC+VX
6	5	37	3	32	7	1	0	7	9	76	0.04	VOC	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
7	4	28	-1	22	5	2	1	-2	17	73	0.04	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	OC+VX
8	1	11	-3	9	2	0	0	-29	13	76	0.04	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	OC+VX
9	0	3	-1	2	0	0	0	-52	39	68	0.05	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	OC+VX
10	0	1	0	0	0	0	0	-54	21	70	0.04	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	OC+VX
11	0	4	0	3	0	0	0	-51	22	81	0.05	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	OC+VX
12	4	29	4	20	4	1	0	60	103	36	0.07	VOC	?	?	?	?	?	?	?	?	?	?	?	?	?	?	expl
13	17	113	13	76	18	9	8	150	41	18	0.04	VOC	17	17	17	17	17	VOY	VOY	VOY	VOC	?	?	?	?	?	?
14	23	155	6	106	25	17	15	183	34	13	0.15	VOC	18	18	18	18	18	VOY	VOY	VOY	VOC	12	VOY	VOY	VOY	VOY	-
15	50	337	27	195	46	43	38	664	47	2	0.22	VOC	45	45	45	45	45	VOY	VOY	VOY	VOC	49	VOY	VOY	VOY	VOY	-
16	66	438	16	223	53	83	74	866	31	1	0.22	VOC	61	61	61	61	61	VOY	VOY	VOY	VOC	65	VOY	VOY	VOY	VOY	-
17	69	461	3	212	50	110	98	907	24	0	0.19	VOC	62	62	62	62	62	VOY	VOY	VOY	VOC	66	VOY	VOY	VOY	VOY	-
18	97	642	28	218	52	96	85	1274	57	0	0.21	VOC	90	90	90	90	90	VOY	VOY	VOY	VOC	94	VOY	VOY	VOY	VOY	-
19	100	661	3	178	42	99	88	1312	23	1	0.21	VOC	91	91	91	91	91	VOY	VOY	VOY	VOC	95	VOY	VOY	VOY	VOY	-
20	73	484	-27	87	20	72	64	963	25	1	0.21	VOC	min1	min1	min1	min1	min1	VOX	VOY	VOY	VOC	min1	VOX	VOY	VOY	VOY	-
21	46	305	-27	82	19	76	67	605	57	2	0.12	VOC	min2	min2	min2	min2	min2	CS	VOY	VOY	VOC	min2	CS	VOY	VOY	VOY	-
22	60	400	14	140	33	112	100	795	47	2	0.12	VOC	60	60	60	60	60	VOY	VOY	VOY	VOC	60	VOY	VOY	VOY	VOY	-
23	67	444	7	168	40	106	94	883	27	1	0.10	VOC	67	67	67	67	67	VOY	VOY	VOY	VOC	67	VOY	VOY	VOY	VOY	-
24	78	521	11	234	55	88	78	1032	39	1	0.10	VOC	78	78	78	78	78	VOY	VOY	VOY	VOC	78	VOY	VOY	VOY	VOY	-
25	99	656	21	334	79	65	58	1297	39	0	0.12	VOC	99	99	99	99	99	VOY	VOY	VOY	VOC	98	VOY	VOY	VOY	VOY	-
26	93	617	-6	345	82	49	43	1214	24	1	0.12	VOC	100	100	100	100	100	VOY	VOY	VOY	VOC	99	VOY	VOY	VOY	VOY	-
27	96	636	3	387	92	12	10	1242	27	0	0.08	VOC	101	101	101	101	101	VOY	VOY	VOY	VOC	100	VOY	VOY	VOY	VOY	-
28	91	603	-5	391	93	18	16	1176	16	1	0.10	VOC	102	102	102	102	102	VOY	VOY	VOY	VOC	101	VOY	VOY	VOY	VOY	-
29	89	589	-2	418	100	35	31	1158	23	1	0.08	VOC	min1	min1	min1	min1	min1	VOX	VOY	VOY	VOC	min1	VOX	VOY	VOY	VOY	-
30	67	444	-22	341	81	35	31	868	20	1	0.06	VOC	67	67	67	67	67	VOY	VOY	VOY	VOC	66	VOY	VOY	VOY	VOY	-
31	62	412	-5	321	76	32	28	804	13	0	0.03	VOC	min1	min1	min1	min1	min1	VOX	VOY	VOY	VOC	min1	VOX	VOY	VOY	VOY	-
32	52	348	-10	271	64	24	21	681	15	2	0.04	VOC	min2	min2	min2	min2	min2	CS	CS	CS	CS	CS	CS	CS	CS	CS	CV
33	42	282	-10	205	49	14	12	549	29	1	0.04	VOC	min2	min2	min2	min2	min2	CS	CS	CS	CS	CS	CS	CS	CS	CS	CV
34	35	234	-7	154	36	16	14	453	27	4	0.04	VOC	min2	min2	min2	min2	min2	CS	CS	CS	CS	CS	CS	CS	CS	CS	CV
35	25	168	-10	110	26	15	13	214	14	6	0.03	VOC	min2	min2	min2	min2	min2	CS	CS	CS	CS	CS	CS	CS	CS	CS	CV
36	18	119	-7	67	16	10	8	187	53	9	0.10	VOC	min2	min2	min2	min2	min2	CS	CS	CS	CS	CS	CS	CS	CS	CS	CV
37	19	127	1	58	13	10	8	233	39	4	0.04	VOC	min2	min2	min2	min2	min2	CS	CS	CS	CS	CS	CS	CS	CS	CS	CV
38	13	86	-6	46	11	6	5	202	36	1	0.03	VOC	min2	min2	min2	min2	min2	CS	CS	CS	CS	CS	CS	CS	CS	CS	CV
39	10	67	-3	28	6	6	5	184	34	3	0.06	VOC	min2	min2	min2	min2	min2	CS	CS	CS	CS	CS	CS	CS	CS	CS	CV
40	7	48	-3	16	3	4	3	182	22	3	0.06	VOC	min2	min2	min2	min2	min2	CS	CS	CS	CS	CS	CS	CS	CS	CS	SIL
41	6	42	-1	23	5	2	1	141	51	8	0.04	VOC	min2	min2	min2	min2	min2	CS	CS	CS	CS	CS	CS	CS	CS	CS	CV
42	7	47	1	26	6	2	1	121	38	12	0.05	VOC	min2	min2	min2	min2	min2	CS	CS	CS	CS	CS	CS	CS	CS	CS	CV
43	14	94	7	58	13	2	1	159	31	24	0.05	VOC	14	14	14	14	14	VOY	VOY	VOY	VOY	VOY	VOY	VOY	VOY	VOY	-
44	17	113	3	49	11	5	4	174	42	16	0.10	VOC	15	15	15	15	15	VOY	VOY	VOY	VOY	VOY	VOY	VOY	VOY	VOY	-
45	30	204	13	106	25	12	10	296	33	13	0.11	VOC	28	28	28	28	28	VOY	VOY	VOY	VOY	VOY	VOY	VOY	VOY	VOY	-
46	33	220	3	149	35	17	15	425	45	18	0.11	VOC	29	29	29	29	29	VOY	VOY	VOY	VOY	VOY	VOY	VOY	VOY	VOY	-
47	38	253	5	187	44	5	4	476	23	14	0.09	VOC	30	30	30	30	30	VOY	VOY	VOY	VOY	VOY	VOY	VOY	VOY	VOY	-
48	22	147	-16	114	27	5	4	187	30	23	0.04	VOC	min1	min1	min1	min1	min1	VOX	VOX	VOX	VOY	VOY	VOY	VOY	VOY	VOY	-
49	16	106	-6	81	19	3	2	157	21	31	0.03	VOC	min2	min2	min2	min2	min2	VOX	VOX	VOX	VOY	VOY	VOY	VOY	VOY	VOY	-
50	11	76	-5	46	11	4	3	135	40	27	0.04	VOC	min2	min2	min2	min2	min2	VOX	VOX	VOX	VOY	VOY	VOY	VOY	VOY	VOY	-

vocaliques (CV), avec deux degrés de certitude dont le plus faible est étiqueté CV ?.

La sortie du superviseur MODE est une séquence de [VOY + VOX] (voyelle), FRIC (bruit d'explosion ou constrictive), CSB (tenue d'occlusive ou silence), CS (consonne constrictive, consonne occlusive voisée ou consonne vocalique), CV (consonne vocalique), CV ? (consonne vocalique putative).

3. Les décisions du superviseur phonétique

Le superviseur phonétique SUPERMODE prend la sortie de l'analyseur acoustico-phonétique MODE et les résultats binaires et flous des processeurs TRAITS2 et TRAITS 3. La sortie de MODE contient toutes les voyelles possibles représentées par les séquences [VOY + VOX], SUPERMODE traite les consonnes.

Le module KL8 recherche les transitions fortes en comparant les deux dérivées DE et DK avec les transitions symboliques du trait d'ouverture fournies par le processeur d'indices flous TRAITS3. Ensuite sont appelés les modules KL9, KL10 et KL11 qui identifient respectivement les bruits d'explosion forte, la structure fine des modes d'articulation et les bruits d'explosion faible. On sait que les bruits d'explosion forte qui caractérisent [*k*] et [*t*] devant les voyelles antérieures sont accompagnés de valeurs élevées de PZ et sont étiquetés FRIC.

Ainsi l'identification correcte de certaines séquences FRIC comme des bruits d'explosion permet également d'émettre des hypothèses sur le lieu d'articulation de la consonne et les traits du contexte vocalique.

Mais l'identification d'un bruit d'explosion sous l'étiquette FRIC ne peut faire appel au seul critère de seuil (valeur de PZ au-dessus d'un seuil somme toute arbitraire). Nous savons par la base de connaissances phonétiques que lorsque FRIC représente un bruit d'explosion, EN et DK se trouvent dans un certain rapport.

La sortie de KL9 identifie sur ces critères les explosions fortes étiquetées EXPL. Plus tard les symboles EXPL seront considérés comme des requêtes pour émettre l'hypothèse de la présence du trait AIGU sur la consonne et la voyelle subséquente.

Le module KL10 interprète les symboles CSB, FRIC et CS. Les CSB sont reconnus soit comme SIL (Silence) soit comme [OC + VX] (occlusive voisée) : le critère de distinction est contenu dans les indices du trait d'ouverture du processeur flou TRAITS3.

Les occlusives non voisées qui pour être identifiées comme telles impliquent la connaissance du contexte ne peuvent être dérivées de l'interprétation des seules séquences CSB.

Les suites FRIC non encore traitées sont étiquetées soit CT-VX (constrictives non voisées) soit CT + VX (constrictives voisées) sur la base d'une comparaison entre la contribution du premier canal (K1) et les indices du trait d'ouverture fournis par TRAITS3.

Les suites CS sont interprétées SIL (silence), CSV (consonne voisée) ou CV (consonne vocalique) ; les critères sont respectivement l'état des indices acoustiques du trait d'ouverture à la sortie de TRAITS3, la comparaison entre EN et la sortie de TRAITS3, la comparaison entre EN et les indices binaires du trait d'acuité à la sortie de TRAITS2. Enfin, KL11, à l'aide de connaissances contextuelles, recherche les bruits d'explosion faible ; deux degrés d'explosion faible sont définis au-dessous de EXPL : EXPL2 et **expl.**

Nous présentons, à titre d'exemple, deux des règles de KL11 :

1. {if((avant = = « SIL ») && ($14 = = « CT-VX »))
   {$14 = « EXP 2 » ; print}}

2. {if((avant = = « SIL ») && ($13 = = « ? »))
   {$14 = « **expl** », print}}

La règle 1 se lit : si le champ 14 est étiqueté CT-VX (constrictif non voisé) et si cet échantillon est précédé de l'étiquette SIL (silence), alors l'échantillon CT-VX pointé dans le champ 14 se réécrit EXP2 (explosion de niveau 2).

La règle 2 identifie, selon le même critère contextuel, **expl** (explosion faible) sous le symbole « ? ».

Le module KL11 doit être considéré comme une récursivité de KL9 après que nous ayons obtenu l'information nécessaire sur le mode d'articulation des séquences.

4. Discussion

La stratégie qui sous-tend le modèle SAPHO est une analyse des propriétés qui définissent les classes acoustiques et phonétiques et qui permettent de passer des classes fondamentales, voyelles et consonnes, à des segments subphonétiques.

L'analyse de chaque macroclasse est équivalente à une suite de règles ordonnées : ces règles sont des règles indépendantes du contexte, des règles contextuelles et des règles de construction. Nous présentons ci-dessous l'essentiel de ces règles sous leur forme simplifiée (note 2) :

VOYELLES

1. Règles indépendantes du contexte

   MAX → VOY
   MIN1 → VOX

2. Règles dépendantes du contexte

   MIN2 → VOX /— {VOX, CSB, FRIC, SIL}

3. Règles de construction

   [VOY, VOX] → voyelle

CONSONNES

1. Règles indépendantes du contexte

   CSB → {SIL, OC + VX }
   FRIC → {EXPL, FRIC}
   FRIC → {CT – VX, CT + VX }

2. Règles dépendantes du contexte

   min2 → CS/—VOY
   CT – VX → EXP2/SIL —
   ? → expl/ {SIL, OC + VX } —

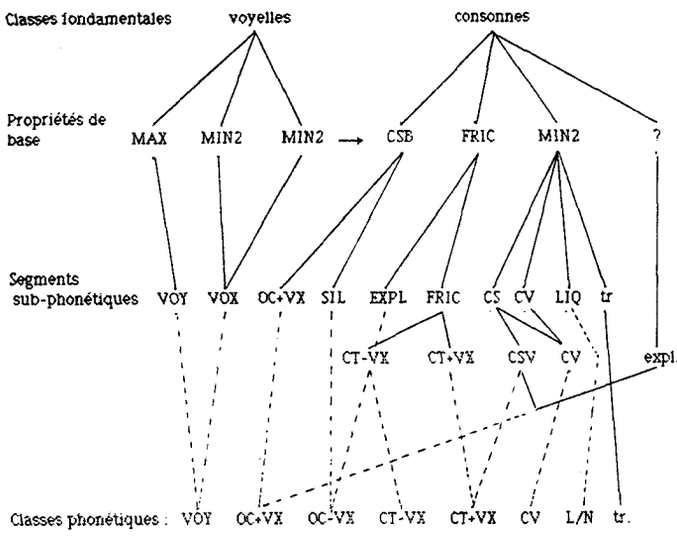
3. Règles indépendantes du contexte

   CS → {CS, CV, LIQ, tr. }
   CS → {CSV, CV }
   LIQ → {L, N}
   CSV → CT + Vx
   CSV → CSV

4. Règles de construction

   SIL + EXP1 → OC – VX1
   SIL + EXP2 → OC – VX2
   CSV + expl. → OC + VX

Ces règles peuvent être représentées par le graphique suivant :



Le dernier niveau du graphe (classes phonétiques) qui n'est pas inclus dans SAPHO est relié aux couches infraphonétiques par des pointillés. Le modèle SAPHO d'autre part ne contient pas encore les règles d'identification des unités phonémiques. Les connaissances phonologiques ne sont pas disponibles à cette étape.

Il va sans dire que le passage au dernier niveau phonétique du graphe n'est pas aussi simple que le laissent supposer la structure du graphe ou les règles simplifiées de construction. Toutefois on aura remarqué que les segments dérivés des séquences étiquetées (couches subphonétiques) sont pertinents pour les unités du niveau phonétique.

Le niveau de représentation phonétique comprend des macroclasses telles que les voyelles, les constrictives, les occlusives etc... qui peuvent se trouver dans une relation de biunivocité avec les unités du niveau phonémique.

La sortie de SAPHO fournit certaines de ces unités phonétiques, mais elles sont organisées comme des allophones et possèdent des caractéristiques qui apparaissent par exemple dans une transcription étroite ; SAPHO fournit donc la structure allophonique interne de ces mêmes unités.

Ainsi nous savons que [*r*] apical, consonne vocalique, possède une organisation interne avec un ou plusieurs battements séparés par des segments vocaliques. Nous trouvons cette organisation chez le locuteur M. R. La consonne [R] postérieure a souvent la même structure temporelle, mais dans certains contextes, devant un silence par exemple, dans les variantes méridionales du français, elle est réalisée comme constrictive. Nous trouvons cet allophone chez le locuteur EG.

De la même manière, l'organisation phonétique de ce qu'on appelle les liquides, est représentée par une discontinuité acoustique autour d'une portion vocalique, comme l'avait montré P. Delattre. C'est cette organisation qu'implique la séquence [LIQ VOY/VOX LIQ] qui devra dans une étape ultérieure être identifiée soit comme une latérale, soit comme une nasale (la notion de *liquide* introduite par les grammairiens grecs incluait justement les latérales et les nasales). D'autres symboles comme **tr** qui indiquent la présence d'une transition forte complètent la structure interne des voyelles ; ces symboles de transition pourront piloter plus tard des hypothèses sur le lieu d'articulation.

4.1. Nous avons dit plus haut que les discussions a priori sur la hiérarchie et la nature des segments obtenus par une procédure de segmentation automatique fondée exclusivement sur un seuillage acoustique ne présentent pas un intérêt majeur, puisque les résultats de la segmentation dépendent du but spécifique de chaque module SA dans les systèmes RAP.

Toutefois, il est utile et intéressant de définir la nature des unités ou des segments que l'on peut obtenir à la sortie d'un système piloté par l'information acoustique et phonétique.

A la première étape,on obtient des propriétés acoustiques (PA) ; des séquences PA homogènes, on déduit des segments acoustiques [4, 11]. L'analyse ultérieure implique une activité pilotée par des requêtes qui supposent l'accès de plus en plus pressant à des connaissances phonétiques. A ce stade, nous obtenons des unités phonétiques allophoniques qui doivent être traitées par une procédure descendante pilotée par les connaissances phonétiques et phonologiques, procédure qui permet d'identifier les unités phonétiques qui se trouvent dans une relation de biunivocité avec les unités phonémiques de haut niveau.

Bien que les premiers niveaux de dérivation fournissent des segments acoustiques nous devons insister sur la nécessité absolue d'utiliser les connaissances phonétiques pour piloter l'extraction des propriétés acoustiques ; c'est seulement de cette manière que les propriétés acoustiques pourront par la suite être utilisées comme des faits susceptibles de piloter les processus de calcul et que ces derniers pourront fournir des segments acoustiques pertinents pour les unités phonétiques.

La hiérarchie proposée, propriétés acoustiques, segments acoustiques, unités phonétiques, est obtenue dans SAPHO par une stratégie dont le point de départ est l'identification du plus petit nombre possible de séquences. Cette stratégie, qui atteint les segments phonétiques par un découpage de plus en plus fin des séquences de départ, permet de réduire au maximum l'indéterminisme au début du processus de reconnaissance. Ainsi, on évite les faux aiguillages, au premier stade de la procédure, par le recours aux indices robustes, la recherche d'îlots de confiance et la segmentation indirecte sur la base des séquences de propriétés acoustico-phonétiques. Nous n'avons pas retenu la segmentation préalable à partir d'un seuil prédéfini de la fonction d'instabilité ; cette procédure en effet gèle prématurément les frontières entre des sous-segments qui sont difficilement pertinents pour les segments phonétiques, puisque dans cette hypothèse, la segmentation n'est pas pilotée par les connaissances phonétiques.

La stratégie que nous avons choisie peut être considérée comme une adaptation de l'algorithme de construction de

niveaux (« Level Building ») utilisé par Meyers et Rabiner pour l'identification de mots dans une phrase « to recognize a sequence of *n* words, at least *n* levels must be built. The first corresponds to hypothesizing that the unknown phrase consists of a simple word, the 2nd level of 2 words hypotheses etc... The results from each level are taken as inputs to the calculations of the next level » [16].

5. Résultats

Les modules de SAPHO ont été testés sur un corpus de 20 mots prononcés par un locuteur homme.

Un test complémentaire a été effectué sur les mêmes mots prononcés par deux autres locuteurs (1 homme, 1 femme). L'évaluation a été menée sur un corpus de 60 mots prononcés par 5 autres locuteurs (3 hommes, 2 femmes).

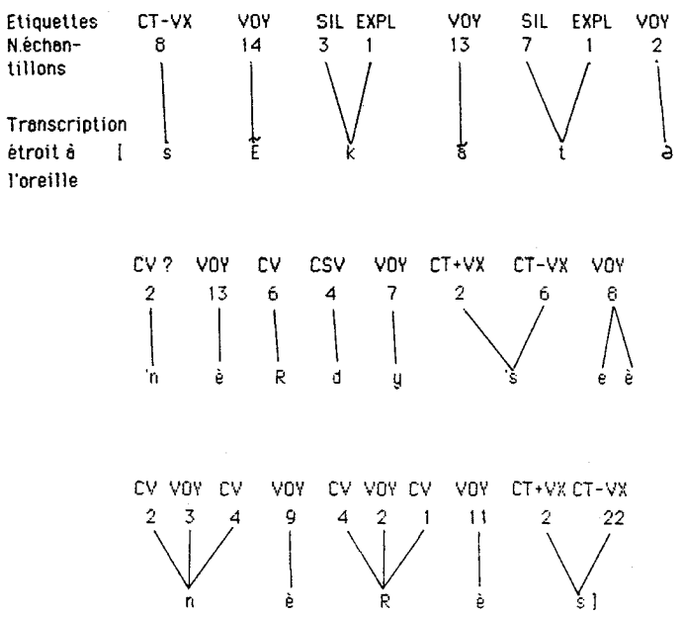
Une évaluation complémentaire est prévue sur un corpus de mots isolés et des phrases de la base de données du français BDSONS.

L'évaluation sur le second corpus de 60 mots donne des résultats encourageants, avec un score de 96,5 %, sur l'identification des consonnes : 1,5 % d'insertions et 2 % de CV omises sur un total de 650 consonnes.

SAPHO ne peut pas encore être évalué sur les groupes consonantiques à cause de leur rareté dans le corpus, bien que l'organisation interne de certains groupes présents soit correctement reconnue (par exemple, dans blague, castor, cal(e)pin, chartreuse...).

SAPHO a été également testé sur quelques phrases. Nous trouverons ci-après le résultat obtenu sur l'énoncé « *Cinquantenaire du CNRS* » prononcé par un septième locuteur (D.A.) :

1) Sortie du système :



Les lignes qui joignent les étiquettes aux unités phonétiques n'indiquent pas des insertions ou des omissions à la sortie de SAPHO, mais unissent les unités phonétiques à leur structure interne. J'ai déjà dit précédemment que SAPHO ne contient pas encore les règles qui permettent la construction des unités phonétiques et phonologiques. On voit par cet exemple la complexité de ce niveau d'interprétation ; en effet, la séquence CV VOY CV pourra être interprétée soit comme une suite phonétique [nèR], soit comme l'organisation interne d'une même consonne vocalique, telle que [n] ou [R]. Ces règles devront faire appel à des contraintes phonotactiques, sans négliger les critères concrets tels que le rapport de durée entre VOY et l'entourage consonantique.

Conclusion

Nous avons présenté et discuté le modèle SAPHO mis en œuvre en langage AWK sous UNIX, sur une station de travail Masscomp. Ce système est conçu comme une procédure de segmentation indépendante du locuteur fondée sur une reconnaissance préalable du mode d'articulation phonétique. Dans la plupart des modèles RAP, les connaissances phonétiques sont toujours utilisées, au moins de façon implicite. Elles doivent l'être de façon explicite. Les unités phonémiques ne peuvent pas être directement construites à partir du signal acoustique ; elles ne sont pas encore disponibles à la sortie de SAPHO. En effet les macroclasses phonémiques (classes de mode d'articulation), pour être identifiées, nécessitent le recours à l'information phonétique concernant le lieu et le mode d'articulation et la connaissance des contraintes phonologiques et lexicales.

Nous avons tenté de montrer comment les paramètres acoustiques doivent être activés par les connaissances phonétiques si l'on veut obtenir des propriétés acoustiques et des segments pertinents pour la structure interne des unités phonétiques.

Suivant le modèle de Construction de Niveaux (*Level Building*), SAPHO fournit un ensemble hiérarchisé de propriétés et de segments acoustiques, de propriétés et de segments phonétiques qui se trouvent dans un rapport de biunivocité avec les unités phonétiques et leur structure interne. La segmentation est le résultat d'une analyse phonétique de l'onde acoustique.

La souplesse de ce système est assurée par sa modularité. Les modules sont conçus soit comme des processeurs activés par des faits avec un résultat numérique en sortie, soit comme des processeurs pilotés par un plan d'action avec une sortie symbolique. La récursivité dans les superviseurs acoustico-phonétique et phonétique à chaque étape de l'analyse assure la vraisemblance des décisions. La fiabilité de SAPHO est corroborée par l'exactitude des résultats.

Remerciements

Je remercie Eric Giraud, boursier BDI CNRS à l'institut de Phonétique d'Aix en Provence, qui a conçu et mis en œuvre sur une station de travail Masscomp le vocodeur qui est utilisé dans ce travail.

Notes

(1) Ce Vocoder a été développé et mis en œuvre sur une station de travail Masscomp par Eric Giraud, boursier BDI du CNRS, à l'Institut de Phonétique d'Aix-en-Provence.

Bandes des canaux du vocodeur en Hz

K1	250	450
K2	450	625
K3	625	875
K4	875	1 050
K5	1 050	1 325
K6	1 325	1 625
K7	1 625	1 875
K8	1 875	2 200
K9	2 200	2 500
K13	3 500	3 875
K14	3 875	4 300
K15	4 300	4 800

(2) [ ] signifie : sommation temporelle des items entre crochets
( ) signifie : choix facultatif des items entre parenthèses
{ } signifie : choix obligatoire d'un des items entre accolades.

Manuscrit reçu le 5 mars 1990.

BIBLIOGRAPHIE

[1] A. V. Aho, B. W. Kerninghan and P. J. Weinberger (1988), *The AWK Programming Language,* Addison Wesley, Amsterdam.

[2] J. Allen (1982), Implementation of Models for Speech Recognition, in *Haton* 1982, pp. 217-230.

[3] A. Bonneau, M. Rossi and G. Mercier (1985), Hierarchical recognition of French vowels by expert system IROISE-SERAC, *Symposium franco-suédois,* Grenoble, avril 1985.

[4] J. Caelen, G. Perennou and N. Vigouroux (1982), Structuration des informations acoustiques dans le projet ARIAL, *Speech Communication,* vol. 2, 2-3, pp. 219-222.

[5] P. Demichelis, R. De Mori and P. Laface (1982), Interaction between Auditory, Syllabic and Lexical Knowledge in a Speech Understanding Systems, in *Haton* 1982, pp. 165-178.

[6] R. De Mori and Y. Ching Suen (1985), *New Systems and Architectures for Automatic Speech Recognition and Synthesis,* Nato ASI Series, Series F, vol. 16, Springer, Berlin.

[7] R. De Mori and P. Laface (1985), On the Use of Phonetic Knowledge for Automatic Speech Recognition, in *De Mori and Ching Suen* 1985, pp. 569-592.

[8] C. Dours, M. De Calmès, H. Kabré, J. M. Pécatte, G. Perennou and N. Vigouroux (1989), A Multi-level Automatic Segmentation System : SAPHO and VERIPHONE, *Eurospeech* 89, vol. II, pp. 83-86.

[9] A. Falaschi (1989), Decodifica acustico-fonetica del messaggio vocale su basi informativo-strutturali, *Tesi di dottorato,* Roma.

[10] F. Fallside and W. A. Woods (1985), *Computer Speech Processing,* Prentice-Hall, London.

[11] L. Fissore, P. Laface, G. Micca and R. Pieraccini (1989), Lexical access to large vocabularies for speech Recognition. *I.E.E.E. Trans. Acoustics Speech and Signal Processing,* 37, 8, 1197-1213.

[12] M. Grenié (1987), *Nature et hiérarchie d'indices acoustiques indépendants du locuteur : application à la R.A.S. des voyelles,* Thèse 3e cycle, Aix-en-Provence.

[13] J. P. Haton (1982), *Automatic Speech Analysis and Recognition,* D. Reidel, Dordrecht, Holland.

[14] W. A. Lea (1982), Selecting, Designing and Using Practical Speech Recognizers, in *Haton* (1982), pp. 331-368.

[15] G. Mercier, M. Gilloux, C. Tarridec and J. Vaissière (1985), A New Rule-Based Expert System for Speech Recognition, in *De Mori and Ching Suen* 1985, pp. 303-342.

[16] C. S. Meyers and L. R. Rabiner (1981), A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition, *I.E.E.E. Trans. Acoustics Speech and Signal Processing,* vol. 29, pp. 284-297.

[17] R. K. Moore (1985), Systems for Isolated and Connected Word Recognition, in *De Mori and Ching Suen* 1985, pp. 73-144.

[18] J. Ohala (1985), Linguistics and Automatic Processing of Speech, in *De Mori and Ching Suen* 1985, pp. 447-476.

[19] M. Rossi (1981), De la physiologie à la perception phonémique, *Modèles Linguistiques,* III, 2, pp. 5-22.

[20] M. Rossi, Y. Nishinuma and G. Mercier (1983), Indices acoustiques multilocuteurs et indépendants du contexte pour la reconnaissance de la parole, *Speech Communication,* 2, pp. 215-217.

[21] M. Rossi, A. Bonneau, Y. Nishinuma and M. Grenié (1991), Automatic Recognition of vowels using acoustic cues (à paraître).

[22] E. J. M. Van Mierlo, E. Blaauw and G. Bloothooft (1989), Phoneme Segmentation of Speech based on Temporal Decomposition using Band Filter Spectra and Phonetic Rules, *Eurospeech 89,* vol. II, pp. 71-74.

[23] A. Waibel, Toshiyuzi Hanazawa, G. Hints, Kihohiro Shikano and K. Y. Lang (1989), Phoneme Recognition using Time-Delay neural networks in *I.E.E.E. Trans. Acoustics Speech and Signal Processing,* 37, 3, pp. 328-339.

[24] V. Zue (1982), Acoustic Phonetic Knowledge Representation : implications from Spectrogram Reading Experiments, in *Haton* 1982, pp. 101-120.