Reconnaissance globale et analytique dans SYMULDEPHO, un SYstème MULti-locuteurs de DEcodage acoustico-PHOnétique

Alain Ghio, Mario Rossi

Reconnaissance globale et analytique dans SYMULDEPHO, un SYstème MULti-locuteurs de DEcodage acoustico-PHOnétique

Alain Ghio et Mario Rossi

Institut de phonétique d'Aix-en-Provence
Laboratoire "Parole et Langage" URA 261, CNRS
29, Av.R.Schuman, 13621 Aix-en-Provence
☎ 42.20.43.56

Résumé

Cette étude s'inscrit dans le cadre de la reconnaissance automatique de la parole et décrit un système multilocuteurs de décodage acoustico-phonétique. Le système que nous proposons s'inspire, dans ses grandes lignes, du traitement cognitif humain dans le sens où le stimulus de parole alimente non pas un, mais différents modules de décodage fonctionnant en parallèle. Ces modules, variant par leur conception, fournissent différentes réponses qu'un moteur d'inférence exploite afin de prendre une décision. Le module de reconnaissance globale, de type métrique, utilise la technique P.L.P. (Prédiction Linéaire sur un spectre Perceptif) comme méthode d'extraction de l'information acoustique et effectue une comparaison spectrale par alignement dynamique temporel (Data Time Warping) entre un stimulus et des références. Le résultat de cette analyse globale est un ensemble de couples Consonne/Voyelle. Le module de reconnaissance analytique par réseaux est basé sur les connaissances de l'expert phonéticien. Les réseaux, qui sont des graphes orientés à transitions d'états, sont construits de telle façon à modéliser toutes les variantes contextuelles d'un phonème. Chaque réseau, spécialisé pour un phonème, est stimulé à priori. La découverte d'un chemin dans le réseau provoque une sortie et entraîne ainsi une proposition de candidat pour la reconnaissance analytique. Le moteur d'inférence n'est pas fonctionnel à l'heure actuelle.

Mots-clés : décodage acoustico-phonétique, reconnaissance globale, reconnaissance analytique, système à base de connaissances

Abstract

The topic of this study is automatic speech recognition and concerns more precisely speaker independent acoustic-phonetic decoding. The model that we present is roughly based on the human cognitive processing system: the speech signal excites several parallel distributed processes of decoding. All these processes are conceptually different and they consequently do not give the same output. The final decision is taken by an "intelligent" control process which studies all these responses. The global recognition process, which is a metric method, extracts acoutsic features using a P.L.P. algorithm (Perceptually based Linear Prediction). Then, the Data Time Warping method (DTW) allows us to spectrally compare stimuli to references. Partial results consist in a set of Consonant/ vowel groups. The analytic recognition process is based on phonetic knowledge. It uses networks which are oriented graphs with state transitions. They are supposed to model all allophones of a phoneme. Each network specialized for a precise phoneme is stimulated without distinction. If a path is found along the network, an output appears at the end and a phoneme candidate is proposed. For the moment, the general control process is not implemented.

Keywords : acoustic-phonetic decoding, global recognition, analytical recognition, knowledge-based system

Introduction

Le phonéticien a pour tâche d'apporter des connaissances sur le fonctionnement du langage et de tester ces connaissances. Les technologies vocales, en particulier la reconnaissance automatique, sont un bon banc d'essai pour cette validation. Cette tâche est d'autant plus pressante que nous vivons une époque où le succès des méthodes stochastiques conduit certains spécialistes de la reconnaissance automatique de la parole -et non des moindres - à dénier toute efficacité, voire toute validité, aux systèmes à base de connaissances. Nous ne rappellerons pas le jugement fameux d'un non moins fameux spécialiste de la question. Ce jugement démontre simplement que les phonéticiens n'ont pas su convaincre par le résultat de leurs recherches, malgré le développement remarquable de celles-ci depuis une quinzaine d'années. Beaucoup reste à faire.

1. Une reconnaissance fondée sur les connaissances

1.1. Connaissance ou probabilisme ?

Le langage et la parole sont structurés à la production, de façon complexe, sur plusieurs niveaux. Ils le sont également à la réception. Pour reconnaître, l'auditeur doit décoder cette structure complexe. Dans cette tâche, les connaissances jouent un rôle fondamental. Ce savoir, qui n'est ni automatique, ni immédiat, nécessite plusieurs années d'acquisition à l'enfant en bas âge ou à l'adulte apprenant. Dans la communication orale homme-machine, le dispositif de décodage de la parole a pour rôle de se substituer à l'auditeur. Dans ce cas-là, à nouveau, nous pensons que les connaissances sont essentielles. Toutefois, même si l'on se situe dans la perspective d'un modèle de perception, les méthodes stochastiques telles que les modèles de Markov et les réseaux de neurones se justifient pleinement. En effet, le cerveau est un système adaptatif et sélectif. La densité de connexions entre neurones et groupes de neurones diversifiés explique cette faculté. Et "c'est l'activité sensori-motrice qui sélectionne les groupes neuronaux donnant la sortie ou les comportements adéquats, ce qui permet d'aboutir à la catégorisation. Dans les systèmes de ce type, les décisions sont fondées sur la statistique des corrélations entre signaux" (Edelman, 1992, p.120). Les modèles connexionnistes copient efficacement ce fonctionnement du cerveau. Mais ils n'en copient que ce qui donne lieu à ce qu'Edelman appelle la conscience primaire. A un autre niveau, dans la conscience d'ordre supérieur, la catégorisation symbolique est intimement liée aux mémoires conceptuelle et symbolique, donc aux connaissances. L'information y est traitée par corrélation, à plusieurs niveaux d'information différents. Il s'agit de ne pas confondre les moyens et les finalités. S'en tenir, dans ces conditions, aux méthodes stochastiques équivaudrait à confondre la mémoire avec les mécanismes nécessaires à son établissement, c'est à dire avec les modifications synaptiques (Edelman, 1992). Le cerveau est un manipulateur de symboles (Millikan, 1984) ; la parole est un échange de symboles dont les valeurs sont fondées sur les connaissances acquises par apprentissage et sur une perpétuelle recatégorisation des stimuli sensoriels variables à l'aide des connaissances. Le fonctionnement du cerveau et du langage fonde ainsi la légitimité des systèmes à base de connaissances. Il reste à démontrer qu'ils sont efficaces : tâche de longue haleine !

1.2. Vers une imitation du traitement cognitif : la parallélisation

Le cortex cérébral est organisé sur un ensemble de "cartes" qui constituent des structures locales stratifiées fortement interconnectées, dont les connexions sont massivement réentrantes (Edelman, 1992). Un même stimulus arrive parallèlement sur plusieurs cartes. La catégorisation provient des connexions réentrantes d'une part entre ces cartes, d'autre part entre ces cartes et les mémoires (Figure 1). La catégorisation est donc le résultat de corrélations entre les informations traitées en parallèle. Le cerveau est un corrélateur. On peut vouloir imiter ce mode de fonctionnement du cerveau, ce qui conduit à la mise en oeuvre de systèmes multi-analyses et multi-experts contrôlés par un moteur d'inférence. Ce type de système est bien connu et développé en particulier au CRIN à Nancy (Carbonnel et al., 1986; Fohr, 1986). Nous nous orientons dans cette même direction.

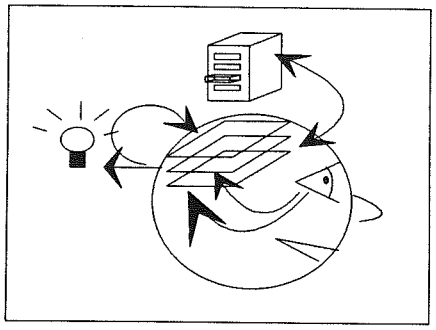


Figure 1 : Parallélisme et catégorisation

1.3. Présentation du système de reconnaissance "SYMULDEPHO"

Le SYstème MULtilocuteurs de DEcodage acoustico-PHOnétique (SYMULDEPHO) que nous présentons est composé de plusieurs modules qui diffèrent à la fois par leurs algorithmes de représentation du signal et par la conception du décodage (Figure 2).

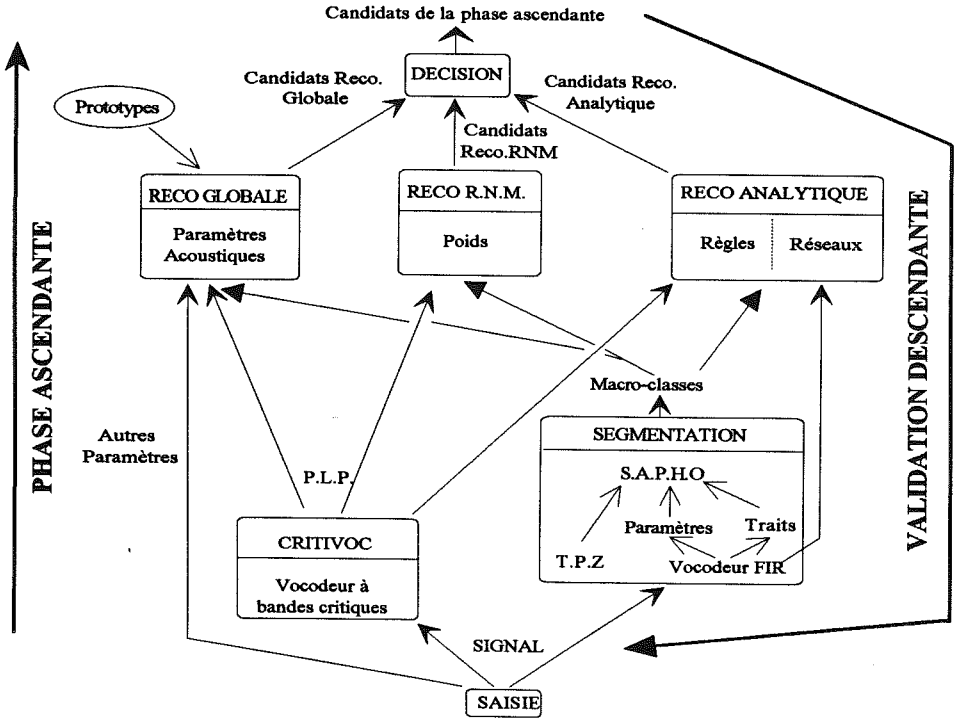


Figure 2 : Schéma fonctionnel de SYMULDEPHO

Le principe est de faire fonctionner ces dispositifs en parallèle, chacun d'entre eux analysant le signal de parole de façon indépendante. Dans une approche multi-modale, on pourrait envisager d'associer en juxtaposition avec le canal oral un décodage du canal visuel (analyse labiale). Une telle démarche permet d'exploiter les redondances de la communication parlée, chaque module de décodage fournissant un résultat conforté ou infirmé par ses voisins. De plus, elle autorise la levée d'ambiguités et la résolution de problèmes parfois non résolus si l'on fonctionne linéairement. Dans tous les cas, notre objectif actuel n'est pas de fournir un système complet et performant à cours terme; il s'inscrit dans un long processus de recherche et d'acquisition de connaissances. Les différents modules de SYMULDEPHO sont les suivants : reconnaissance globale métrique, décodage analytique par réseaux à transition d'états, décodage analytique par règles, reconnaissance par réseaux neuro-mimétiques. De façon générale, le décodage est dynamique et inclut le contexte

avec les effets de la coarticulation. De plus, la catégorisation est prévue pour être le résultat de la vérification de l'information montante par l'information descendante activée dans la mémoire des connaissances (Gilles, 1993). La segmentation préalable du signal de parole est effectuée à l'aide de S.A.P.H.O. (Rossi, 1990), algorithme de segmentation automatique conçu sur la base de règles phonétiques définies par un expert. L'outil d'analyse acoustique est un vocodeur à canaux qui utilise une série de 15 filtres FIR répartis quasi uniformément dans le domaine spectral et agissant directement par convolution sur le signal. Le résultat de la segmentation est un étiquetage des trames d'analyse en macro-classes (ex: voyelle, occlusive voisée, silence...) qui permet un repérage des événements et une analyse orientée. La reconnaissance par réseaux neuro-mimétiques reste encore au stade de projet. L'objectif est d'utiliser un dispositif à apprentissage par rétropropagation de type TDNN où les cellules d'entrée sont alimentées par des coefficients acoustiques PLP et dont la sortie fournit un codage par traits (Rodellar et al., 1991 ; Nishinuma et al., 1993). Nous ne présentons ici que les deux premiers modules de reconnaissance dans l'analyse montante.

2. Le module de reconnaissance globale

2.1. Description de la méthode

Le principe de la reconnaissance globale est d'effectuer une comparaison spectrale entre un stimulus et des références. La méthode choisie pour le module de reconnaissance globale implique trois opérations :

- l'extraction de l'information acoustique
- la comparaison métrique des paramètres du stimulus avec des prototypes
- la prise de décision.

2.1.1. L'extraction de l'information acoustique

'CritiVoc' est un vocodeur qui utilise une modélisation de phénomènes psycho-acoustiques (pondération sonique, intégration par bandes critiques) et effectue une représentation compacte temps / fréquence du signal de parole. Le traitement auditif effectué par 'CritiVoc' (Figure 3) consiste, tout d'abord, à corriger le signal d'un point de vue spectral en appliquant une pondération sonique sur le spectre brut pour tenir compte de la non-linéarité fréquentielle de l'oreille. Dans un second temps, on réalise une intégration du spectre par bandes critiques (Zwicker & Therhart., 1980). Cette répartition spectrale à bandes est ensuite interpolée puis modélisée par une prédiction linéaire. La prédiction linéaire a l'avantage de modéliser le spectre auditif dans son ensemble en fournissant un codage par pôles (Boite & Kunt, 1987).

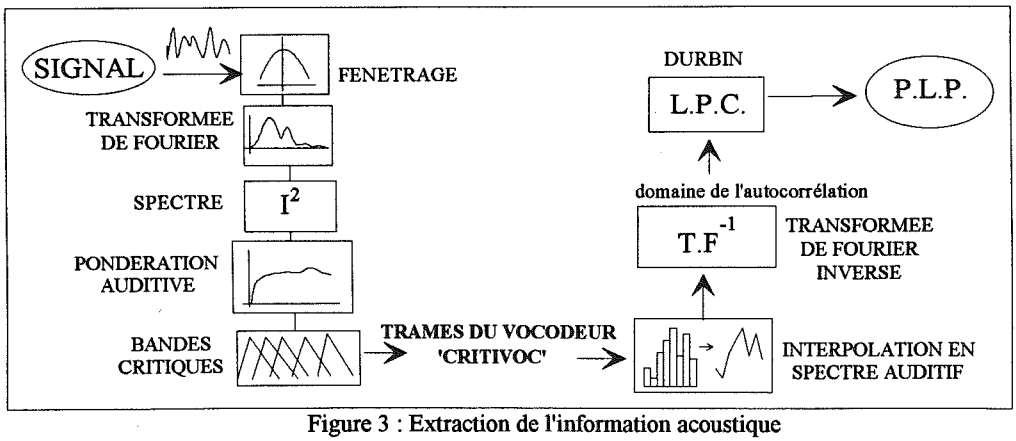


Figure 3 : Extraction de l'information acoustique

Le bilan de l'opération est le suivant: si 'N' est le nombre de trames d'analyse sur le signal et 'M' est l'ordre de la prédiction linéaire, l'extraction nous fournit N séries de M paramètres acoustiques appelés coefficients de Prédiction Linéaire sur un spectre Perceptif ou encore

coefficients PLP (Hermansky, 1990 ; Junqua, 1990). La Figure 4 illustre le traitement mis en place au cours de cette analyse. On y distingue 4 grandes étapes : le calcul du spectre par Transformée de Fourier (T.F.R), la correction spectrale par application d'une pondération sonique, l'intégration par bandes critiques et la modélisation par Prédiction Linéaire (PLP).

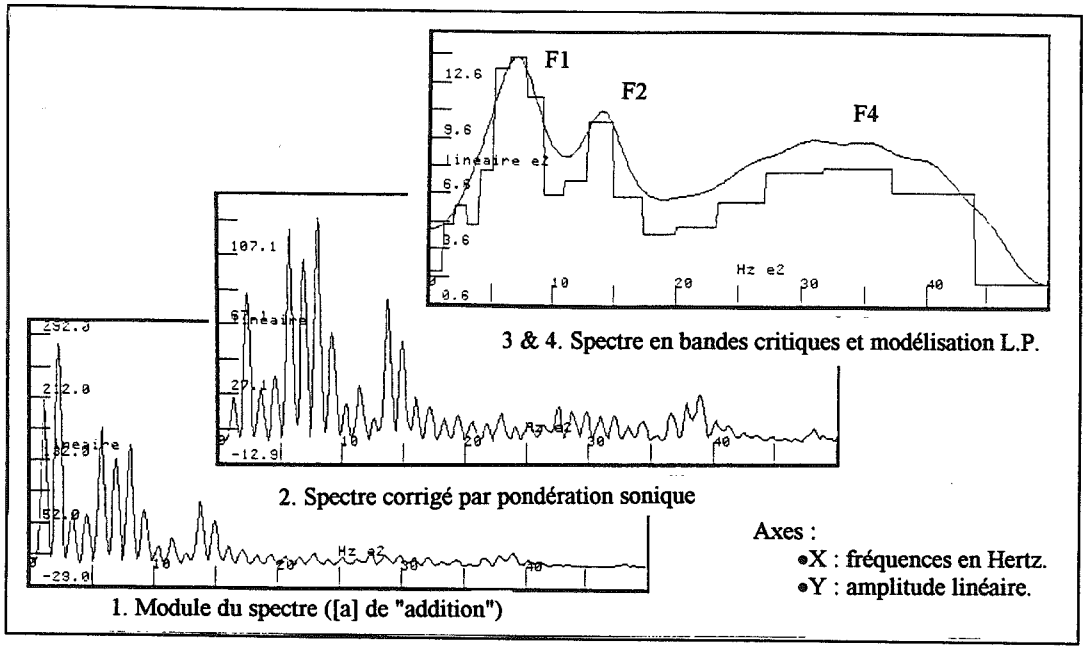


Figure 4 : Exemple du traitement P.L.P pour le [a] de "addition"
(1. Spectre d'amplitude, 2. Spectre pondéré, 3. Intégration en bandes critiques 4. Modèle PLP)

2.1.2. La comparaison métrique

Le principe de la reconnaissance globale est d'effectuer une comparaison spectrale entre un stimulus et des références acoustiques afin d'estimer un degré de similitude. Dans notre module de reconnaissance globale, cette opération est réalisée en calculant une distance entre les coefficients PLP issus du signal à reconnaître et ceux qui sont relatifs à chacune des références que l'on a archivées préalablement. La technique d'alignement dynamique temporel (Data Time Warping mieux connu sous le nom de D.T.W.) est une méthode très répandue depuis longtemps (Vintsyuk, 1968 ; Haton, 1974). Elle a pour but d'ajuster les échelles temporelles de deux éléments à comparer. Cette transformation non linéaire permet ainsi de synchroniser des segments acoustiques de même nature entre le stimulus et la référence. Ainsi, dans le cas de couples Consonne/Voyelle, on effectue une comparaison globale du diphone mais la partie consonantique du stimulus est mise en correspondance avec la partie consonantique du prototype, de même pour les parties vocaliques. La conséquence est l'alignement des transitions C/V. La technique DTW établit la distance cumulée entre le stimulus et chacun des prototypes. Chacune de ces valeurs nous donne ainsi un degré de similitude entre deux réalisations acoustiques. Ces valeurs nous permettent d'émettre des hypothèses sur les prétendants possibles, la distance minimale correspondant au prétendant le plus sérieux. Toutefois, compte tenu de la variabilité, on ne conserve pas que le meilleur candidat. La supervision de la liste des prétendants est nécessaire avant de prendre une décision.

2.1.3. La prise de décision

Dans le développement actuel du système, la reconnaissance s'effectue sur des diphones de type Consonne/Voyelle. Le vocabulaire se compose des combinaisons entre les 6 occlusives [p,t,k,b,d,g] et 7 voyelles [a,i,u,o,e,y,oe]. Pour chaque couple Consonne/Voyelle, on possède 10 prototypes en archive. Au cours de la reconnaissance, le système calcule les coefficients PLP du

stimulus à reconnaître et les compare à ceux des prototypes. L'identification du couple CV se fonde sur les 20 premiers candidats issus de cette comparaison.

2.1.3.1. Identification de la voyelle

Le processus de reconnaissance de la voyelle suppose, dans un premier temps, la prise en compte du nombre d'apparitions de chaque voyelle parmi ces 20 prétendants. Les voyelles retenues sont celles qui apparaissent au moins 5 fois parmi les 20. On a ainsi un nombre variable de voyelles candidates (de une à trois, deux étant le chiffre le plus fréquent). Dans un second temps, on prend en considération les valeurs des distances associées à ces voyelles retenues. La voyelle déclarée la plus probable sera celle qui aura les distances les plus faibles...

2.1.3.2. Identification de la consonne

Une analyse globale comme celle que nous utilisons ne nous permet pas une reconnaissance immédiate des consonnes occlusives. Cette impossibilité vient du fait que d'un point de vue phonétique, cette classe de phonèmes est extrêmement délicate à analyser (très grande variabilité, évènements acoustiques rapides). Dans une démarche analytique, le décodage s'effectue par la recherche de traits phonétiques du niveau acoustique vers les niveaux symboliques. La solution envisagée pour la reconnaissance globale est inverse. On laisse le module de comparaison fournir une liste de candidats CV. A partir de ces prétendants, on évalue les traits les plus fréquents. Ainsi, si parmi les 10 premiers candidats, on en trouve 8 dont la consonne est sourde, on peut se permettre d'exclure tous les diphones dont la consonne est voisée. On procède ainsi par éliminations successives.

2.1.4. Bilan

La sortie du module de reconnaissance globale fournit donc une liste de couples CV prétendants, auxquels un trait de confiance est attaché compte tenu du caractère marqué ou non des traits de la consonne et de la valeur de la distance associée à la voyelle.

2.2. Le réglage des paramètres d'analyse

La performance du système repose sur le choix judicieux des caractéristiques d'analyse. Celles-ci doivent être assez générales pour passer l'obstacle de la variabilité mais aussi suffisamment précises pour conserver un pouvoir discriminant. L'étape d'extraction des caractéristiques acoustiques du signal nous semble essentielle. La recherche de la meilleure configuration se traduit par un calcul de taux d'erreurs de reconnaissance en fonction de diverses valeurs de paramètres. Une première estimation du choix des paramètres a été effectuée grâce aux études antérieures et à une évaluation qualitative (étude graphique des résultats de la modélisation PLP susceptible de donner une configuration initiale convenable). La démarche qui a suivi a consisté à étudier systématiquement l'évolution des performances du système en faisant varier les paramètres. Cette étude ne sera pas détaillée ici ; nous n'en donnerons que les conclusions. Il s'avère que la meilleure configuration des paramètres d'extraction PLP est la suivante :

- fenêtre de Hamming de longueur égale à 20ms avec un recouvrement de 50%
- pré-emphase à partir de 500Hz et pondération sonique de type X
- spectre d'amplitude en dB
- ordre de prédiction égal à 8

2.3. Les résultats

Les tests ont été effectués sur des diphones de type occlusive + [a, i, u, o, e, y, oe] extraits de logatomes du corpus acoustique SYL de BD-SON (Cervantes et al., 1986). Dix locuteurs sont utilisés, soit un total de 6*7*10 = 420 énoncés. Un étiquetage manuel nous permet de faire porter

l'évaluation uniquement sur le module de reconnaissance globale sans introduire d'erreur de segmentation.

2.3.1. Reconnaissance de la voyelle (Tableau 1 et 2)

On distingue trois types de résultats :

- le cas où la voyelle a été reconnue correctement en première position (*1er candidat)*
- le cas où la voyelle a été reconnue correctement en 2, 3, 4ième position (*2nd rang)*
- le cas où la voyelle n'a pas été reconnue (*Omis*sion)

Tableau 1 : Résultats de la reconnaissance globale des voyelles

	1er candidat	2nd rang	Omission
/a/	95%	3%	2%
/i/	95%	3%	2%
/u/	90%	8%	2%
/o/	83%	17%	0%
/e/	70%	25%	5%
/y/	77%	15%	8%
/oe/	93%	5%	2%
Moyenne	86%	11%	3%

Le tableau 2 représente la matrice de confusion des voyelles. On remarque que, bien souvent, dans les cas de mauvaise reconnaissance, le système se reporte sur un voisin acoustique du stimulus : /o/ est confondu avec /a/ et /u/, tous trois graves ; /y/ est confondu avec /e/ et /i/, tous trois aigus. La confusion générale avec /oe/ s'explique par la position centrale de ce phonème dans l'espace acoustique. La réalisation en [oe] est souvent le résultat d'une neutralisation d'autres phonèmes due à des effets de contexte. La confusion est donc facile..

Tableau 2 : Matrice de confusion des voyelles pour la reconnaissance globale

réponse / stimulus	/a/	/i/	/u/	/o/	/e/	/y/	/oe/
/a/ =>	***95%***	0%	0%	0%	0%	0%	5%
/i/ =>	0%	***95%***	0%	0%	2%	3%	0%
/u/ =>	0%	0%	***90%***	8%	2%	0%	0%
/o/ =>	3%	0%	12%	***83%***	0%	0%	2%
/e/ =>	0%	5%	0%	0%	***70%***	8%	17%
/y/ =>	0%	2%	0%	0%	8%	***77%***	13%
/oe/ =>	0%	0%	0%	0%	5%	2%	***93%***

2.3.2. Reconnaissance de la consonne (Tableau 3)

Dans le cas de la reconnaissance des consonnes, le système fournit plusieurs prétendants dont aucun n'est privilégié à priori. Il n'existe pas de classement parmi les candidats. On distingue alors deux types de résultats : le cas où la consonne stimulus se trouve parmi les candidats (*Correct),* le cas où la consonne stimulus ne se trouve pas parmi les candidats (*Omission).* La colonne *'Nb moyen de candidats'* indique le nombre moyen de candidats fournis par le système dans le décodage de chacune des consonnes. Ce paramètre est important car intuitivement, on comprend que plus ce nombre sera grand, moins bonne sera la discrimination mais meilleurs seront les résultats bruts.

Tableau 3 : Résultats de la reconnaissance globale des consonnes

	Correct	Omission	Nb moyen de candidats
/p/	96%	4%	1.86
/t/	97%	3%	2.07
/k/	96%	4%	1.89
/b/	97%	3%	2.10
/d/	99%	1%	2.24
/g/	97%	3%	2.13
Moyenne	97%	3%	2.05

L'analyse succinte des résultats est la suivante: en moyenne, le dispositif de reconnaissance globale propose 2 occlusives parmi les 6 possibles ; dans 97% des cas, la bonne réponse est parmi ces 2 candidats.

2.3.3. Bilan de la reconnaissance globale

Afin d'estimer la pertinence de l'analyse PLP, nous avons effectué des tests comparatifs en utilisant d'autres types de paramètres acoustiques tels que ceux issus de la technique LPC, du Cepstre, du Mel Cepstre (Davis & Mermelstein, 1980 ; Brancaccio et al., 1992). Nous avons conservé le même module de comparaison DTW ainsi que celui de la décision. Les résultats seront présentés dans une publication ultérieure. Il semble, d'après ces tests, que la technique PLP ait un degré de pertinence analogue à la méthode MelCepstre. Les erreurs de décodage en utilisant l'une ou l'autre des techniques n'étant pas les mêmes, nous envisageons d'utiliser en parallèle les deux types d'extraction. De plus, la technique PLP est loin d'être figée et nous gardons l'espoir d'améliorer son pouvoir discriminant en tenant compte des causes responsables des mauvais résultats du décodage, entre autres de /e/ et /y/. L'exemple de /y/, pour laquelle F2 et F3 sont confondus dans un seul pôle autour de 1700-2000Hz, montre ici les limites de la reconnaissance globale. Enfin, il faut garder à l'esprit que les résultats présentés ici ne sont relatifs qu'à un décodage sur l'axe paradigmatique, le découpage temporel étant fourni par un étiquetage manuel. Tout risque d'omission ou d'insertion sont évités. L'étape d'évaluation globale du système reste à faire en tenant compte d'une méthodologie précise comme le propose Bourjot et al. (1990).

3. Le module de reconnaissance analytique par réseaux

Habituellement on conçoit des réseaux syntagmatiques munis de probabilités (modèles de Markov). Notre but n'est pas d'obtenir un score élevé à tout prix, mais, comme il est indiqué dans l'introduction, de tester des connaissances ; c'est à dire de rechercher les connaissances qui sont censées être utilisées par l'auditeur et qui fournissent les meilleurs scores. Les réseaux utilisés ne seront donc pas munis de probabilités (probabilités qui, certes, sont très efficaces, mais qui masquent une partie de notre ignorance). Deuxième différence avec les modèles de Markov : les réseaux conçus ici pour la reconnaissance des voyelles fonctionnent sur l'axe paradigmatique à chaque trame du signal à la sortie du vocodeur.

3.1. Les données

Les fichiers de données contiennent les répartitions spectrales obtenues toutes les 10ms par un vocodeur à canaux qui utilise 15 filtres FIR agissant sur le signal par convolution temporelle. La contribution de chaque bande à l'organisation acoustique est représentée par le pourcentage de l'énergie dans chaque canal par rapport à l'énergie totale (Figure 5). Le programme de reconnaissance des voyelles VOYLIEU fait appel au module de segmentation SUPERMODE, au programme TRAIT-FLOU qui extrait un certain nombre d'indices de chaque trame du signal, aux réseaux de reconnaissance des voyelles LieuA, LieuO, LieuY, etc...

	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10	K11	K12	K13	K14	K15	EN
1	10	0	0	0	0	0	0	18	34	27	9	0	0	0	0	1
2	2	1	1	2	8	1	11	6	25	23	3	2	4	5	0	20
3	3	2	1	3	6	2	10	13	20	17	5	3	4	3	1	62
4	32	4	3	1	5	3	7	9	10	10	2	4	2	2	0	111
5	54	9	1	1	2	3	4	12	2	3	1	1	0	0	0	155
6	72	5	1	1	1	0	2	8	3	1	0	0	0	0	0	292
7	59	16	3	0	1	0	1	6	6	1	0	0	0	0	0	363
8	57	13	3	0	0	0	0	4	15	1	0	0	0	0	0	460
9	53	9	3	0	0	0	1	4	21	1	0	0	0	0	0	493
10	47	8	3	0	0	0	2	3	28	1	0	0	0	0	0	546
11	41	11	3	0	0	1	1	3	32	1	0	0	0	0	1	606
12	33	13	3	0	1	1	1	10	28	3	0	0	1	0	1	613
13	32	14	4	1	2	1	2	13	20	3	0	0	0	0	0	514
14	33	9	7	5	7	4	2	14	9	3	0	1	0	0	0	307
15	29	4	10	9	9	5	7	7	6	5	1	1	0	0	0	170
16	26	3	4	5	9	5	15	10	5	6	1	2	1	0	1	92
17	11	2	15	0	16	2	0	6	6	14	1	3	0	0	0	71
18	7	3	17	9	24	2	4	9	4	14	0	1	0	0	0	91
19	12	11	21	10	20	2	5	2	4	8	0	0	0	0	0	165
20	17	25	17	4	11	2	8	3	5	3	0	0	0	0	0	329
21	21	30	12	2	3	1	13	5	6	2	0	0	0	0	0	523
22	25	40	9	0	0	2	5	2	6	2	0	0	1	0	0	568
23	23	47	10	0	2	3	1	0	4	1	0	0	1	0	0	621
24	18	53	9	0	6	3	0	0	2	1	0	0	1	0	0	717
25	14	55	8	0	10	1	0	0	1	4	0	0	0	0	0	830
26	12	57	8	0	14	0	0	0	0	2	0	0	0	0	0	902
27	13	62	9	1	7	0	0	0	0	1	0	0	0	0	0	851
28	17	59	12	1	3	0	0	0	0	1	0	0	0	1	0	786
29	18	54	13	1	4	0	0	0	0	1	0	0	0	2	0	775
30	21	52	11	4	2	0	0	0	0	1	0	0	0	3	0	846
31	21	58	7	6	1	0	0	0	0	0	0	0	0	1	0	925
32	19	65	4	5	0	0	0	0	0	0	0	0	0	0	0	1115
33	17	70	4	4	0	0	0	0	0	0	0	0	0	0	0	1242
34	13	66	8	6	0	0	0	0	0	0	0	0	0	0	0	1290
35	12	44	14	23	0	0	0	0	0	0	0	0	0	0	0	1193
36	19	23	16	36	1	0	0	0	0	0	0	0	0	0	0	997
37	27	10	10	47	1	0	0	0	0	0	0	0	0	0	0	1003
38	41	16	9	27	1	0	0	0	0	0	0	0	0	0	0	876
39	36	20	27	12	1	0	0	0	0	0	0	0	0	0	0	869
40	22	18	50	4	1	0	0	0	0	0	0	0	0	0	0	782
41	11	11	64	7	3	0	0	0	0	0	0	0	0	0	0	576
42	4	17	61	9	6	0	0	0	0	0	0	0	0	0	0	290

Figure 5 : Répartition spectrale sur le mot "culot", locuteur C (Femme). Analyse centiseconde par filtres FIR (voir dans le texte). Les valeurs numériques dans chaque bande représentent la contribution de chaque canal par rapport à l'énergie totale (somme des énergies des canaux).

3.1.1. Le module 'SUPERMODE'

SUPERMODE est l'algorithme de segmentation de SAPHO (Rossi, 1991). La segmentation indirecte est déduite des connaissances. A la sortie de SUPERMODE, on obtient un signal analysé et étiqueté. Les étiquettes réfèrent aux macro-classes de mode d'articulation : voyelle, consonne vocalique, constrictive,occlusive, etc...(Figure 6).

N	EN%	EN	DE	BF	BF%	HF	HF%	ZER	DK	PZ	RZ	K3	TRO	K3	TRO	K3	K6	TRO	K3	K4	K5	K6	K8	MOD
1	0	1	0	0	0	0	0	-39	0	0.52	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	SIL
2	1	20	1	1	0	13	5	133	74	0.38	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	CSB	SIL
3	4	62	3	5	0	40	15	182	31	0.11	VOC	?	?	?	?	?	?	?	?	?	?	?	?	expl
4	8	111	4	44	3	43	16	146	69	0.06	VOC	5	5	5	5	5	VOY	VOY	VOY	VOY	VOY	VOY	VOY	-
5	12	155	4	100	8	29	11	164	61	0.05	VOC	6	6	6	6	6	VOY	VOY	VOY	VOY	VOY	VOY	VOY	-
6	22	292	10	230	19	35	13	554	37	0.05	VOC	16	16	16	16	16	VOY	VOY	VOY	VOY	VOY	VOY	VOY	-
7	28	363	6	283	23	47	18	691	33	0.06	VOC	17	17	17	17	17	VOY	VOY	VOY	VOY	VOY	VOY	VOY	-
8	35	460	7	335	27	92	35	875	18	0.05	VOC	24	24	24	24	24	VOY	VOY	VOY	VOY	VOY	VOY	VOY	-
9	38	493	3	320	26	128	49	946	15	0.05	VOC	25	25	25	25	25	VOY	VOY	VOY	VOY	VOY	VOY	VOY	-
10	42	546	4	316	26	174	67	1052	16	0.06	VOC	26	26	26	26	26	VOY	VOY	VOY	VOY	VOY	VOY	VOY	-
11	46	606	4	333	27	218	84	1182	16	0.05	VOC	27	27	27	27	27	VOY	VOY	VOY	VOY	VOY	VOY	VOY	-
12	47	613	1	300	25	257	100	1206	25	0.05	VOC	28	28	28	28	28	VOY	VOY	VOY	VOY	VOY	VOY	VOY	-
13	39	514	-8	262	21	185	71	1003	19	0.05	VOC	min1	min1	min1	min1	min1	VOY	VOY	VOY	VOY	VOY	VOY	VOY	-
14	23	307	-16	165	13	82	31	594	34	0.04	VOC	23	23	23	23	23	VOY	VOY	VOY	VOY	VOY	VOY	VOY	-
15	13	170	-10	88	7	34	13	155	37	0.05	VOC	min1	min1	min1	min1	min1	VOY	VOY	VOY	VOY	VOY	VOY	VOY	-
16	7	92	-6	34	2	23	8	137	30	0.05	VOC	min2	min2	min2	min2	min2	CS	CS	CS	CS	CS	CS	CS	CSV
17	5	71	-2	25	2	21	8	116	63	0.05	VOC	min2	min2	min2	min2	min2	CS	CS	CS	CS	CS	CS	CS	CV
18	7	91	2	32	2	25	9	147	28	0.05	VOC	min2	min2	min2	min2	min2	CS	CS	CS	CS	CS	CS	CS	CV
19	12	165	5	89	7	23	8	200	37	0.05	VOC	min2	min2	min2	min2	min2	CS	CS	CS	CS	CS	CS	CS	CV
20	25	329	13	207	17	36	14	693	48	0.05	VOC	25	25	25	25	25	VOY	VOC	?	?	?	?	?	?
21	40	523	15	339	28	67	26	1021	34	0.06	VOC	40	40	40	40	40	VOY	VOC	40	VOY	VOY	VOY	VOY	-
22	44	568	4	420	35	62	24	1106	35	0.05	VOC	41	41	41	41	41	VOY	VOC	41	VOY	VOY	VOY	VOY	-
23	48	621	4	486	41	37	14	1212	22	0.05	VOC	42	42	42	42	42	VOY	VOC	42	VOY	VOY	VOY	VOY	-
24	55	717	7	573	47	28	10	1399	19	0.07	VOC	49	49	49	49	49	VOY	VOC	49	VOY	VOY	VOY	VOY	-
25	64	830	9	639	53	41	15	1620	18	0.08	VOC	58	58	58	58	58	VOY	VOC	58	VOY	VOY	VOY	VOY	-
26	69	902	5	694	57	18	7	1754	12	0.07	VOC	59	59	59	59	59	VOY	VOC	59	VOY	VOY	VOY	VOY	-
27	65	851	-4	723	60	8	3	1657	16	0.06	VOC	60	60	60	60	60	VOY	VOC	60	VOY	VOY	VOY	VOY	-
28	60	786	-5	699	58	15	5	1532	15	0.05	VOC	min1	min1	min1	min1	min1	VOX	VOC	min1	VOX	VOY	VOY	VOY	-
29	60	775	0	674	56	23	8	1510	10	0.06	VOC	min1	min1	min1	min1	min1	VOX	VOC	59	VOY	VOY	VOY	VOY	-
30	65	846	5	744	62	33	12	1652	12	0.11	VOC	min1	min1	min1	min1	min1	VOX	VOC	60	VOY	VOY	VOY	VOY	-
31	71	925	6	851	70	9	3	1805	16	0.11	VOC	71	71	71	71	71	VOY	VOC	61	VOY	VOY	VOY	VOY	-
32	86	1115	15	1036	86	0	0	2175	15	0.12	VOC	86	86	86	86	86	VOY	VOC	76	VOY	VOY	VOY	VOY	-
33	96	1242	10	1179	98	0	0	2429	8	0.11	VOC	96	96	96	96	96	VOY	VOC	86	VOY	VOY	VOY	VOY	-
34	100	1290	4	1199	100	0	0	2525	14	0.12	VOC	97	97	97	97	97	VOY	VOC	87	VOY	VOY	VOY	VOY	-
35	92	1193	-8	1109	92	0	0	2331	46	0.14	VOC	min1	min1	min1	min1	min1	VOX	VOC	min1	VOX	VOX	VOX	VOX	-
36	77	997	-15	937	78	0	0	1944	44	0.10	VOC	min1	min1	min1	min1	min1	VOX	VOC	min1	VOX	VOX	VOX	VOX	-
37	77	1003	0	942	78	0	0	1956	38	0.09	VOC	min1	min1	min1	min1	min1	VOX	VOC	min1	VOX	VOX	VOX	VOX	-
38	67	876	-10	814	67	0	0	1702	41	0.09	VOC	67	67	67	67	67	VOX	VOC	67	VOX	VOX	VOX	?tr	-
39	67	869	0	825	68	0	0	1688	42	0.09	VOC	min1	min1	min1	min1	min1	VOX	VOC	min1	VOX	VOX	VOX	tr	tr
40	60	782	-7	735	61	0	0	1514	47	0.08	VOC	min1	min1	min1	min1	min1	VOX	VOC	59	VOY	?	?	?	?
41	44	576	-16	535	44	0	0	1102	37	0.08	VOC	min2	min2	min2	min2	min2	VOX	VOC	min1	VOX	VOX	VOX	tr	tr
42	22	290	-22	263	21	0	0	530	21	0.08	VOC	min2	min2	min2	min2	min2	VOX	VOC	min2	VOX	VOX	VOX	tr	tr

Figure 6 : Résultat de la segmentation et de la reconnaissance des macroclasses par SAPHO. Champs 2 à 11 : paramètres ; champs 12 à 25 : étapes de la segmentation (voir Rossi, 1991). Mot **"culot"**, locuteur C (Femme).

3.1.2. Le module TRAIT-FLOU

Dans le programme TRAIT-FLOU, les tests définissent 9 indices : 3 pour le trait d'ouverture / fermeture, 6 pour le trait d'acuité /gravité. 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 compte tenu de certaines conditions, chaque trait est représenté par un symbole sur trois échelles (trois canaux ou sommes de canaux du vocodeur) à quatre degrés (12 degrés au total pour chaque trait, Figure 7) :

Echelle 1 = Fermé, degrés F0, F1, F2, F3.
Echelle 2 = Moyen, degrés F4, etc...
Echelle 3 = 0uvert, degrés O4, etc...

Le trait d'ouverture / fermeture est évalué par l'analyse des trois premiers canaux (250-875Hz), la polarité Grave du trait d'acuité est évaluée par celle des canaux 4 à 6 (875-1625 Hz) et la polarité Aiguë de ce même trait par celle des canaux 7 à 13 inclus. (1625-3875 Hz). Chaque maximum d'énergie est pondéré par son degré d'émergence. On obtient ainsi une structure spectrale catégorisée sur des valeurs floues. La structure spectrale symbolique (Figure 7) est obtenue à la suite de l'interprétation des maxima catégorisés sur la base des connaissances phonétiques.

F2	-	-	-	-	-	0	C9	0
F0	-	-	-	-	-	0	C9	0
F0	-	-	-	-	-	0	C9	0
F1	-	-	-	-	-	0	B9	0
F1	-	-	-	-	-	B8	0	0
F1	-	-	-	-	-	A8	0	0
F1	-	-	-	-	-	A8	A9	0
F1	-	-	-	-	-	0	B9	0
F1	-	-	-	-	-	0	B9	0
F1	-	-	-	-	-	0	B9	0
F1	-	-	-	-	-	0	C9	0
F1	-	-	-	-	-	0	C9	0
F2	-	-	-	-	-	0	B9	0
F1	-	-	-	-	-	B8	0	0
F1	-	-	-	-	-	B7	0	0
F1	-	-	-	-	-	B7	0	0
F3	-	06	-	G4	-	B7	B10	0
F3	-	07	-	G4	-	B8	B10	0
F3	-	06	-	G4	-	0	B10	0
-	F4	-	-	G4	-	B7	0	0
-	F4	-	-	-	-	B7	A9	0
-	F4	-	-	-	-	0	A9	0
-	F4	-	-	-	-	0	0	0
-	F4	-	-	-	-	0	0	0
-	F4	-	-	G5	-	0	0	0
-	F4	-	-	G4	-	0	0	0
-	F4	-	-	-	-	0	0	0
-	F4	-	-	-	-	0	0	0
-	F4	-	-	-	-	0	0	0
-	F4	-	-	-	-	0	0	0
-	F4	-	-	-	-	0	0	0
-	F4	-	-	-	-	0	0	0
-	F4	-	-	-	-	0	0	0
-	05	-	G2	-	-	0	0	0
-	F4	-	G1	-	-	0	0	0
F2	-	-	G1	-	-	0	0	0
F1	-	-	G2	-	-	0	0	0
F2	-	-	G2	-	-	0	0	0
F3	-	07	-	-	-	0	0	0
-	-	07	-	-	-	0	0	0
-	-	07	-	-	-	0	0	0

Figure 7 : Spectre catégorisé à la sortie de TRAITFLOU. Champs 1 à 3 (canaux 1 à 3): étiquetage du trait d'ouverture / Champs 4 à 6 (canaux 4 à 6) : étiquetage du trait de gravité / Champs 7 à 9 (canaux 7 à 13) : étiquetage du trait d'acuité. Mot "culot", locuteur C (Femme).

3.1.3. Le module TRAIT-BIN

L'information apportée par TRAIT-FLOU est complétée par la recherche du canal d'énergie maximale (champ 11, Figure 8) et par les résultats de TRAIT-BIN (champs 12 à 20, Figure 8). TRAIT-BIN a été commenté auparavant (Rossi, 1991). Contrairement aux tests de TRAIT-FLOU, ceux de TRAIT-BIN identifient des indices fondés sur la distribution significative de l'énergie dans

des canaux qui ne sont pas forcément contigus. Par exemple, pour le trait de fermeture FERME-0, on effectue le test :

si	[k1 > (k3 + k4)]	où kx = le xième canal du vocodeur.
alors	Ferme-0 = '+'	
sinon	Ferme-0 = '-'	

Dans TRAIT-BIN, pour un trait donné, par exemple OUVERTURE, chaque indice est représenté par sa polarité. Ainsi le codage des quatre indices du trait d'ouverture / fermeture peut prendre la forme - - + + (champ 12, Figure 8). Cette combinatoire d'indices, présente l'avantage de saisir des structures. Les indices sont hiérarchisés sur la base des connaissances phonétiques et dotés d'un coefficient. La combinatoire des indices fournit ainsi une valeur numérique pour le trait présumé: par exemple, 16, pour la combinatoire - - + +, qui apparaîtra sur les voyelles les plus fermées (champ 13, Figure 8). Cette valeur est l'image d'une organisation complexe qu'il est difficile d'appréhender directement et qui complète la structure de type binaire.

F2	-	-	-	-	-	0	C9	0	0	9	--++	16	+++	1	-++	3	----	14	-BEM.	SIL
F0	-	-	-	-	-	0	C9	0	0	9	--+-	10	+++	1	-+-	4	+-++	6	-BEM.	SIL
F0	-	-	-	-	-	0	C9	0	0	9	--+-	10	+++	1	-++	3	+-++	6	-BEM.	exp!
F1	-	-	-	-	-	0	B9	0	0	9	--++	16	+++	1	-++	3	+-+-	6	-BEM.	-
F1	-	-	-	-	-	88	0	0	0	8	--++	16	+++	1	+++	5	+++-	3	+BEM.	-
F1	-	-	-	-	-	A8	0	0	0	8	--++	16	+++	1	++-	6	----	14	+BEM.	-
F1	-	-	-	-	-	A8	A9	0	0	8	--++	16	+++	1	-+-	4	--+-	13	+BEM.	-
F1	-	-	-	-	-	0	B9	0	0	9	--++	16	+++	1	-+-	4	----	14	+BEM.	-
F1	-	-	-	-	-	0	B9	0	0	9	--++	16	+++	1	-+-	4	----	14	+BEM.	-
F1	-	-	-	-	-	0	B9	0	0	9	--++	16	+++	1	-+-	4	----	14	+BEM.	-
F1	-	-	-	-	-	0	C9	0	0	9	--++	16	+++	1	-+-	4	-+--	11	+BEM.	-
F1	-	-	-	-	-	0	C9	0	0	9	--++	16	+++	1	-+-	4	--+-	13	+BEM.	-
F2	-	-	-	-	-	0	B9	0	0	9	--++	16	+++	1	-+-	4	--+-	13	+BEM.	-
F1	-	-	-	-	-	88	0	0	0	8	--++	16	+++	2	+++	5	--+-	13	+BEM.	-
F1	-	-	-	-	-	B7	0	0	0	4	--++	16	---	8	+++	5	----	14	-BEM.	-
F1	-	-	-	-	-	B7	0	0	0	7	--++	16	+-+	2	+++	5	+-++	6	+BEM.	CSV
F3	-	06	-	G4	-	B7	B10	0	0	5	++--	6	---	8	-++	3	--+-	13	-BEM.	CV
F3	-	07	-	G4	-	88	B10	0	0	5	++--	6	---	8	-++	3	--+-	13	-BEM.	CV
F3	-	06	-	G4	-	0	B10	0	0	5	+---	1	---	8	-+-	4	--+-	13	-BEM.	CV
-	F4	-	-	G4	-	B7	0	0	0	5	-+--	8	---	8	++-	6	--+-	13	+BEM.	?
-	F4	-	-	-	-	B7	A9	0	0	7	-+-+	9	+-+	2	++-	6	--+-	13	+BEM.	-
-	F4	-	-	-	-	0	A9	0	0	9	-+-+	9	--+	7	-+-	4	-+--	11	+BEM.	-
-	F4	-	-	-	-	0	0	0	0	0	-+-+	9	---	8	-+-	4	-++-	12	-BEM.	-
-	F4	-	-	-	-	0	0	0	0	5	-+-+	9	---	8	-+-	4	--+-	13	-BEM.	-
-	F4	-	-	G5	-	0	0	0	0	5	-+-+	9	---	8	-+-	4	+-+-	5	-BEM.	-
-	F4	-	-	G4	-	0	0	0	0	5	-+-+	9	---	8	-+-	4	+-+-	5	-BEM.	-
-	F4	-	-	-	-	0	0	0	0	5	-+-+	9	---	8	-+-	4	--+-	13	-BEM.	-
-	F4	-	-	-	-	0	0	0	0	0	-+-+	9	---	8	---	1	--+-	13	-BEM.	-
-	F4	-	-	-	-	0	0	0	0	0	-+-+	9	---	8	---	1	--+-	13	-BEM.	-
-	F4	-	-	-	-	0	0	0	0	0	-+-+	9	---	8	---	1	----	14	-BEM.	-
-	F4	-	-	-	-	0	0	0	0	4	-+-+	9	---	8	---	1	----	14	-BEM.	-
-	F4	-	-	-	-	0	0	0	0	0	-+-+	9	---	8	---	1	----+	10	-BEM.	-
-	F4	-	-	-	-	0	0	0	0	0	-+-+	9	---	8	---	1	----	14	-BEM.	-
-	F4	-	-	-	-	0	0	0	0	4	-+--	8	---	8	---	1	----	14	-BEM.	-
-	05	-	G2	-	-	0	0	0	0	4	-+--	8	---	8	---	1	----+	10	-BEM.	-
-	F4	-	G1	-	-	0	0	0	0	4	-+--	8	---	8	---	1	----+	10	-BEM.	-
F2	-	-	G1	-	-	0	0	0	0	4	--+-	10	---	8	---	1	----+	10	-BEM.	-
F1	-	-	G2	-	-	0	0	0	0	4	--++	16	---	8	---	1	----+	10	-BEM.	-
F2	-	-	G2	-	-	0	0	0	0	4	--++	10	---	8	---	1	----	14	-BEM.	tr
F3	-	07	-	-	-	0	0	0	0	0	++--	6	---	8	---	1	----	14	-BEM.	?
-	-	07	-	-	-	0	0	0	0	4	++--	6	---	8	---	1	----	14	-BEM.	tr
-	-	07	-	-	-	0	0	0	0	4	++--	6	---	8	---	1	----	14	-BEM.	tr

Figure 8 : Champs 1 à 10 : spectre catégorisé / Champ 11 : canaux d'énergie maximale à chaque trame / Champs 12 à 20 : sortie de TRAITBIN / Champ 21 : résultat de la segmentation. Mot **"culot"**, locuteur C (Femme).

3.2. Conception des réseaux

Les réseaux utilisent essentiellement l'information à la sortie de TRAIT-FLOU (répartition spectrale symbolique). Cette information est complétée par la connaissance du canal d'énergie maximale et la sortie de TRAIT-BIN. Le réseau est un graphe orienté à transitions d'états : G = (N,A), où N est l'ensemble des noeuds et A l'ensemble des arcs. Les noeuds du graphe sont alimentés par :

- le contenu symbolique des canaux significatifs d'une voyelle et de ses allophones (champs 1 à 10, Figure 8) .
- éventuellement le numéro du canal d'énergie maximale (champ 11, Figure 8)
- la valeur scalaire correspondant à la combinatoire des indices (champs 13, 15, 17, 19, Figure 8).

Les arcs s(a) = (N1,N2) qui relient deux noeuds matérialisent les transitions d'état et les changements intervenus lors d'une transition (Figure 9). Dans l'état Nx, la tête de lecture - écriture lit un symbole d'entrée e et écrit un symbole de sortie s qui le remplace. Le symbole s est une interprétation phonétique de e opérée sur la base des connaissances relatives à la structure de la voyelle considérée. Une transition de Nx à Nx+1 est alors représentée par t = (e, s). L'ensemble des

symboles s sur un chemin définit les traits de l'allophone reconnu. Le réseau est un automate fini A défini par le quintuplet A = (E, S, N, f, g) où

E = {e1, e2,... em} - m symboles d'entrée
S = {s1, s2,... sn } - n symboles de sortie
N= {N1, N2,... Np } - p états internes
f = N x E ==> N - fonction de transition
g = N x E ==> S - fonction de sortie

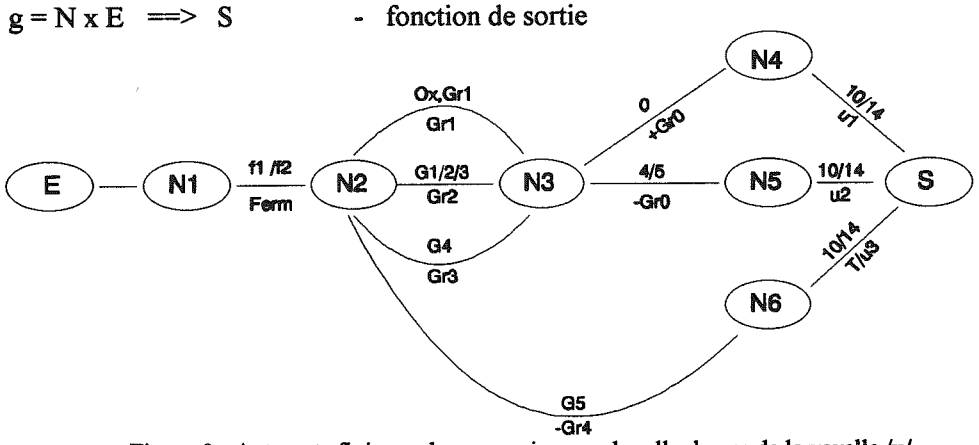


Figure 9 : Automate fini pour la reconnaissance des allophones de la voyelle /u/.

Les fonctions de transition et de sortie peuvent être réunies dans la même table fonctionnement. Celle du réseau LieuOU (Figure 9) a la structure suivante:

Tableau 4 : table de fonctionnement du réseau LieuOU

N \ E	F1/F2	Ox	G1/G2/G3	G4	G5	0	4/5	10/14
N1	N2,Ferm							
N2		N3,Gr1	N3,Gr2	N3,Gr3	N6,-Gra			
N3						N4,+Gr0	N5,-Gr0	
N4								s, u1
N5								s, u2
N6								s, T/u3

Les symboles de sortie fournissent les traits et l'étiquette de l'allophone à reconnaître,à cette restriction près : si un même trait apparaît plusieurs fois avec une valuation identique ou différente, on considère que le dernier trait est une réécriture du précédent et efface ce dernier. Ainsi dans le graphe LieuOU et la table de fonctionnement correspondante, si l'on suit le chemin suivant :

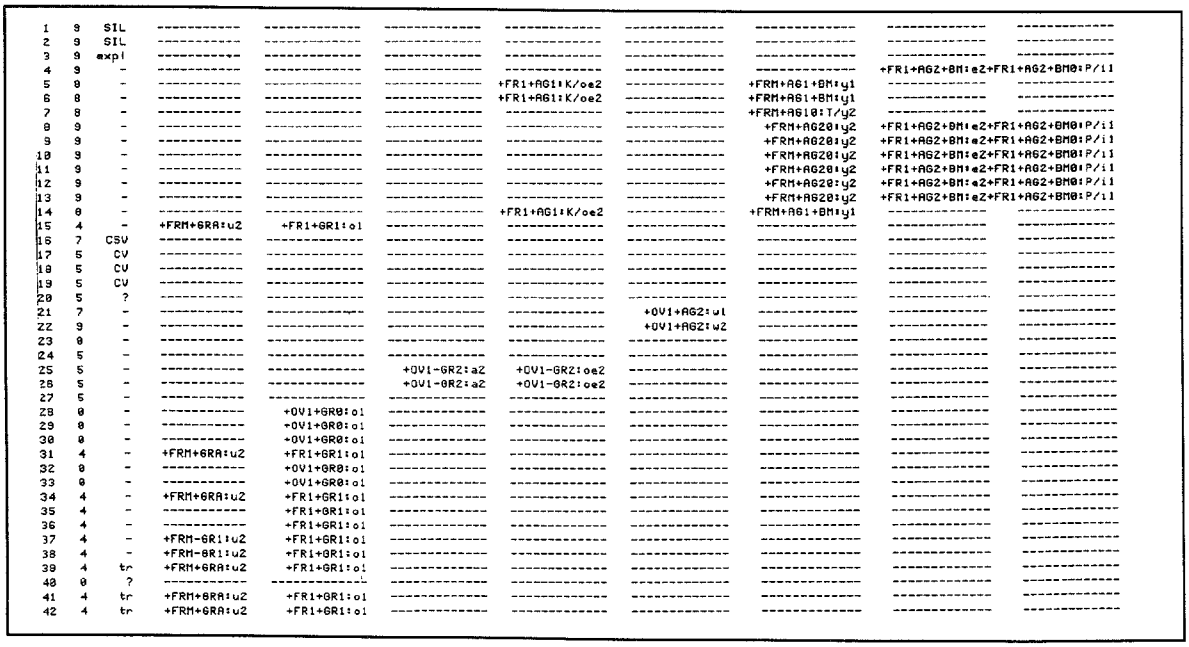
nx	e	s	nx+1
N1	f1/f2	ferm	N2
N2	Ox	gr1	N3
N3	0	gr0	N4
N4	10/14	u1	S

Le trait GR1 qui définit un degré de gravité acoustique est interprété phonétiquement comme +GR0 (+ Grave.0) après la lecture du symbole 0 en N3 (le symbole 0, sur le champ 11 des maxima spectraux spécifie qu'aucun maximum n'est rencontré au-delà du premier formant). En revanche, après la lecture de G5 dans l'état N2,la tête d'écriture imprime la sortie -GRA (- Grave); -GRA n'est pas réécrit.Le symbole de sortie qui définit cet allophone dans l'état N6 est T/u3. L'allophone T/u3

est la variante du phonème /u/ dans le contexte dento-alvéolaire.Un réseau est censé contenir tous les allophones d'un phonème; le réseau ainsi conçu, dont la programmation est simple, est une représentation compacte d'une classe d'équivalence.

3.3. Résultats

Chaque réseau passe sur chaque trame du signal. Tout chemin défini sur un réseau et découvert dans les données provoque une sortie. Plusieurs réseaux peuvent alors répondre sur la même trame (Figure 10). Dans ces conditions plusieurs candidats sont susceptibles d'être proposés pour une même voyelle. Ne sont retenus, dans l'analyse ultérieure, que les candidats qui occupent 70% des trames centrées, c'est à dire des trames qui occupent la partie centrale de la voyelle diminuée de 30% aux bornes, dans la zone de coarticulation. Les candidats retenus ne sont pas classés par ordre ; comme on l'a vu plus haut, aucun critère statistique n'intervient dans le choix. Le critère de décision qui sera mis en oeuvre ultérieurement s'apparente à un raisonnement de type cognitif. Prenons un exemple : pour la voyelle /y/ dans "culot" (Figure 10) sont proposés trois candidats : [y2], [e2] et [p/i1]. [e2] et [i1] sont bémolisés, c'est à dire qu'ils ne présentent aucun maximum d'énergie dans les fréquences supérieures à 2200 Hz ; ces allophones se rencontrent normalement dans un contexte Grave au contact de [p] par exemple, comme cela apparaît dans l'étiquette de l'allophone [p/i1]. Si, sur la base d'autres informations, indépendantes de celles-ci, on est conduit à émettre l'hypothèse que le trait Grave est absent du contexte, alors il sera légitime de classer [y2] comme premier candidat plausible. C'est à cette étape que la sortie des réseaux et leur interprétation sera confrontée à la sortie de la reconnaissance globale de type métrique et à la catégorisation par les RNM. Les allophones proposés par les réseaux sont des variantes contextuelles ; le choix du bon candidat suppose que l'information contextuelle ainsi fournie soit compatible avec le contexte reconnu sur des critères différents. La multiplicité des candidats apporte une information contextuelle qui, grâce à une rétroaction avec le contexte autorise le choix du meilleur candidat. La multiplicité des candidats présente un autre avantage : elle peut jouer un rôle dans la découverte du contexte. Soit l'exemple de la voyelle /o/ de "culot" (Figure 10). Le seul candidat à la sortie des réseaux est [o1] qui est la variante dont le degré de gravité l'apparente à /u/. Mais, à la borne gauche de la voyelle, les réseaux proposent les candidats [a2], [oe2] et [w1], ce qui signifie que le contexte gauche possède le trait Aigu ; si [o1] est la bonne voyelle, le contexte doit correspondre à une consonne dento-alvéolaire (en l'occurrence, il s'agit de la consonne /l/).



1	9	SIL	----------	----------	----------	----------	----------	----------	----------	----------
2	9	SIL	----------	----------	----------	----------	----------	----------	----------	----------
3	9	expl	----------	----------	----------	----------	----------	----------	----------	----------
4	9	-	----------	----------	----------	----------	----------	----------	+FR1+AG2+BM:e2+FR1+AG2+BM0:P/i1	
5	8	-	----------	----------	----------	+FR1+AG1:K/oe2	----------	+FRM+AG1+BM:y1	----------	----------
6	8	-	----------	----------	----------	+FR1+AG1:K/oe2	----------	+FRM+AG1+BM:y1	----------	----------
7	8	-	----------	----------	----------	----------	----------	+FRM+AG10:T/y2	----------	----------
8	9	-	----------	----------	----------	----------	----------	+FRM+AG20:y2	+FR1+AG2+BM:e2+FR1+AG2+BM0:P/i1	
9	9	-	----------	----------	----------	----------	----------	+FRM+AG20:y2	+FR1+AG2+BM:e2+FR1+AG2+BM0:P/i1	
10	9	-	----------	----------	----------	----------	----------	+FRM+AG20:y2	+FR1+AG2+BM:e2+FR1+AG2+BM0:P/i1	
11	9	-	----------	----------	----------	----------	----------	+FRM+AG20:y2	+FR1+AG2+BM:e2+FR1+AG2+BM0:P/i1	
12	9	-	----------	----------	----------	----------	----------	+FRM+AG20:y2	+FR1+AG2+BM:e2+FR1+AG2+BM0:P/i1	
13	9	-	----------	----------	----------	----------	----------	+FRM+AG20:y2	+FR1+AG2+BM:e2+FR1+AG2+BM0:P/i1	
14	8	-	----------	----------	----------	+FR1+AG1:K/oe2	----------	+FRM+AG1+BM:y1	----------	----------
15	4	-	+FRM+GRA:u2	+FR1+GR1:o1	----------	----------	----------	----------	----------	----------
16	7	CSV	----------	----------	----------	----------	----------	----------	----------	----------
17	5	CV	----------	----------	----------	----------	----------	----------	----------	----------
18	5	CV	----------	----------	----------	----------	----------	----------	----------	----------
19	5	CV	----------	----------	----------	----------	----------	----------	----------	----------
20	5	?	----------	----------	----------	----------	----------	----------	----------	----------
21	7	-	----------	----------	----------	----------	+OV1+AG2:u1	----------	----------	----------
22	9	-	----------	----------	----------	----------	+OV1+AG2:w2	----------	----------	----------
23	8	-	----------	----------	----------	----------	----------	----------	----------	----------
24	5	-	----------	----------	----------	----------	----------	----------	----------	----------
25	5	-	----------	----------	+OV1-GR2:a2	+OV1-GR2:oe2	----------	----------	----------	----------
26	5	-	----------	----------	+OV1-GR2:a2	+OV1-GR2:oe2	----------	----------	----------	----------
27	5	-	----------	----------	----------	----------	----------	----------	----------	----------
28	8	-	----------	+OV1+GR0:o1	----------	----------	----------	----------	----------	----------
29	8	-	----------	+OV1+GR0:o1	----------	----------	----------	----------	----------	----------
30	8	-	----------	+OV1+GR0:o1	----------	----------	----------	----------	----------	----------
31	4	-	+FRM+GRA:u2	+FR1+GR1:o1	----------	----------	----------	----------	----------	----------
32	8	-	----------	+OV1+GR0:o1	----------	----------	----------	----------	----------	----------
33	8	-	----------	+OV1+GR0:o1	----------	----------	----------	----------	----------	----------
34	4	-	+FRM+GRA:u2	+FR1+GR1:o1	----------	----------	----------	----------	----------	----------
35	4	-	----------	+FR1+GR1:o1	----------	----------	----------	----------	----------	----------
36	4	-	----------	+FR1+GR1:o1	----------	----------	----------	----------	----------	----------
37	4	-	+FRM-GR1:u2	+FR1+GR1:o1	----------	----------	----------	----------	----------	----------
38	4	-	+FRM-GR1:u2	+FR1+GR1:o1	----------	----------	----------	----------	----------	----------
39	4	tr	+FRM+GRA:u2	+FR1+GR1:o1	----------	----------	----------	----------	----------	----------
40	8	?	----------	----------	----------	----------	----------	----------	----------	----------
41	4	tr	+FRM+GRA:u2	+FR1+GR1:o1	----------	----------	----------	----------	----------	----------
42	4	tr	+FRM+GRA:u2	+FR1+GR1:o1	----------	----------	----------	----------	----------	----------

Figure 10 : Sortie des réseaux sur le mot "culot" (loc. féminin).Le dernier champ est la voyelle ouverte è de degré 1.

L'information contextuelle fournie par les réseaux peut également permettre de restituer une consonne qui n'a pas été identifiée lors de la phase de segmentation. Soit le mot "culot" (locuteur S., Figure 11), les allophones reconnus pour les deux voyelles sont respectivement [y0] et [o1]. Etant donné les contraintes phonologiques sur le lexique, qui excluent la suite yo, l'une des hypothèses plausibles est la présence d'une consonne, entre les allophones [y0] et [o1]; aucune consonne n'avait été identifiée lors de la phase de segmentation.

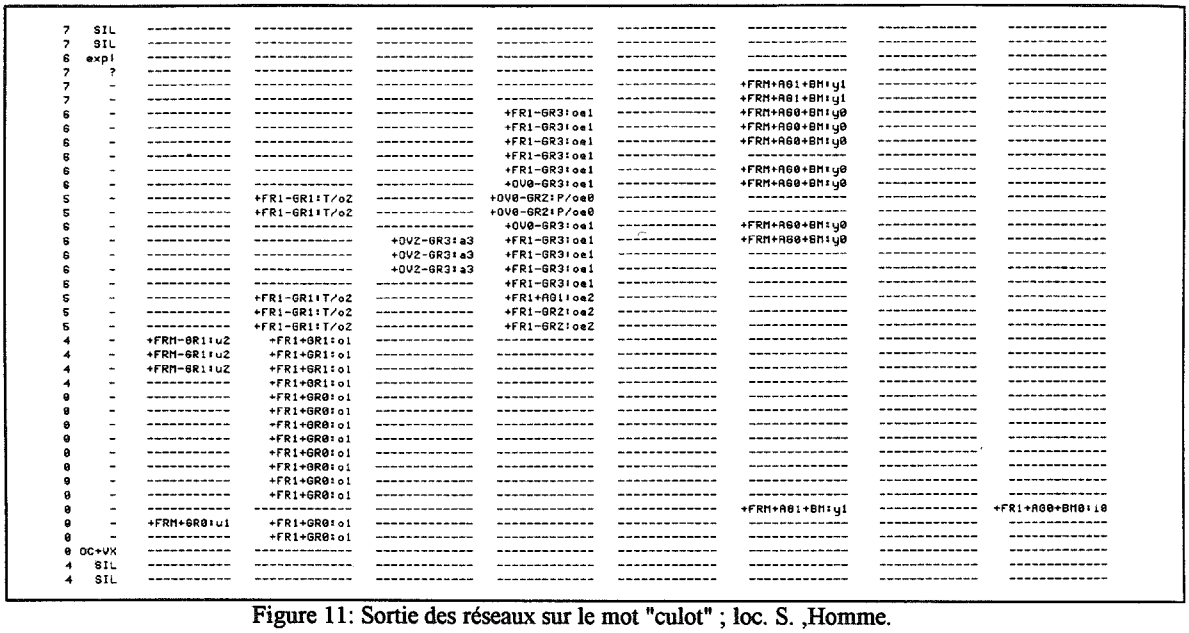


Figure 11: Sortie des réseaux sur le mot "culot" ; loc. S. ,Homme.

Les réseaux, tels qu'ils sont conçus ici, apportent une information contextuelle fondée sur les connaissances, qui est essentielle non seulement dans le choix du bon candidat voyelle, mais également dans la restitution ou l'interprétation du contexte. Cette partie interprétative est en cours de développement. La sortie brute des réseaux a été testée sur un vocabulaire de 42 mots dits par 5 locuteurs (3 hommes et 2 femmes) ; sur un total de 452 voyelles, le bon candidat est présent dans 92% des cas. La sortie des réseaux propose en moyenne 2 candidats par phonème :

1 candidat dans 41% des cas
2 " " 49% " "
3 " " 8% " "

Conclusion

Nous avons présenté un système de DAP multilocuteurs. Ce système est fondé sur le principe des analyses multiples indépendantes ; il prend en compte différentes stratégies qui ont été ou qui sont celles de la reconnaissance automatique aujourd'hui :

- méthode globale de type métrique.
- méthode analytique à base de connaissances.
- méthode stochastique (RNM)

La sortie des différents modules sera traitée par le moteur d'inférence qui sera élaboré ultérieurement. Comme on l'a montré, la sortie des modules autorise un certain nombre d'hypothèses limitées par des contraintes contextuelles. Ces hypothèses conduisent à une vérification descendante à la fois dans le signal et dans le traitement symbolique intermédiaire des modules. Cette démarche, qui s'apparente au raisonnement cognitif, est tout à fait adaptée à la reconnaissance de la parole spontanée où les réalisations acoustiques sont rarement finies et bien formées et où les ambiguïtés qui en découlent ne peuvent être levées que par le contexte et les connaissances des niveaux supérieurs.

Bibliographie

BOITE R., KUNT M. (1987). *Traitement de la parole*. Presses polytechniques romandes.

BOURJOT C., BOYER A., FOHR D., HATON J.P. (1990). *Méthodologies pour l'évaluation phonétique*. Actes des 18ièmes J.E.P, Montréal, 201-206

BRANCACCIO A., CEGLIE F., D'ACUNZO G., PELEAZ C., RICCIO A., RIGOSI F. (1992). *A comparative study of the influence of parameter processing on two different approaches for speech recognition in adverse environment*. Proceedings of 'Speech processing in adverse conditions' . ESCA Conference, Cannes. Université de Nice - Sophia Antipolis. 93-96

CARBONNEL N., HATON J.P., FOHR D., LONCHAMP F., PIERREL J.M. (1986). *APHODEX, design and implementation of an acoustic-phonetic decoding expert system*. Proceedings of IEEE ICASSP, Tokyo, Vol.2, 1201-1204.

CERVANTES O., SERIGNAT J.F., DESCOUT R., CARRE R. (1986). *Définition et réalisation d'une base de données des sons du français*. Actes des 15ièmes JEP, GALF, Aix-en-Provence, 213-216.

DAVIS S.B., MERMELSTEIN P. (1980). *Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences*. Revue IEEE, Vol.ASSP 28, 357-366.

EDELMAN G.M (1992). *Biologie de la conscience*, Ed. O.Jacob, Paris.

FOHR D. (1986). *APHODEX : un système expert en décodage acoustico-phonétique de la parole continue*. Thèse de doctorat, Université de Nancy I.

GILLES P. (1993). *Décodage acoustico-phonétique de la parole et adaptation au locuteur*. Thèse de doctorat, Université d'Avignon.

HATON J.P. (1974). *Contribution à l'analyse, la paramétrisation et la reconnaissance automatique de la parole*. Thèse de doctorat, Université de Nancy I.

HERMANSKY H. (1990). *Perceptual linear predictive (PLP) analysis of speech*. J.Acous.Soc.Am., 87, 1738-1752.

JUNQUA J.C. (1990). *Utilisation d'un modèle d'audition et de connaissances phonétiques en reconnaissance automatique de la parole*. Revue Traitement du signal - GRETSI - numéro spécial, 275-284.

MILLIKAN R.G. (1984). *Language, thought and other biological categories : new foundations for realism*. MIT Press, Cambridge.

NISHINUMA Y., KITASAWA S., SHINMURA T. (1993). *Réseau neuromimétique identifiant les traits distinctifs des voyelles du japonais*. Travaux de l'Institut de phonétique d'Aix, Vol.15, 185-214

RODELLAR V., NAHARRO F., GARCIA C., MARTIN S., MUNOZ M.L., GOMEZ P. (1991). A Neural Network for the extraction and characterization of the phonetic features of speech. *Proceedings of International Conference on Neural Networks & their Applications*, Nîmes.

ROSSI, M. (1990). *Segmentation automatique de la parole : pourquoi ? Quel segments ?* Revue Traitement du signal - GRETSI - numéro spécial , 315-326.

VINTSYUK T.K. (1968) *Speech recognition by dynamic programming*. Kybernetika, 4, 81-88.

ZWICKER E., THERHART E. (1980). *Analytical expression for critical bands rate and critical bandwidth as a function of frequency*. J.Acous.Soc.Am., 68, 1523-1525.