Décodage acoustico-phonétique : problèmes et éléments de solution (*)

Acoustic-phonetic decoding of speech : problems and solutions

Jean-Paul HATON
CRIN/INRIA Lorraine
B.P. 239
54506 Vandœuvre-Lès-Nancy
France

Jean-Paul Haton, agrégé de l'université, est professeur à l'Université de Nancy I où il enseigne divers aspects de l'informatique et de l'intelligence artificielle. Il est responsable, au sein du Centre de Recherche en Informatique de Nancy, de l'équipe *Reconnaissance des Formes et Intelligence Artificielle.* Son principal domaine d'intérêt concerne, depuis plus de quinze ans, la communication homme-machine et l'intelligence artificielle : reconnaissance de la parole, reconnaissance de caractères et de graphismes, traitement et interprétation d'images, systèmes experts et systèmes à bases de connaissances.
Jean-Paul Haton est directeur du GRECO-PRC « Communication Homme-Machine ». Il est actuellement détaché à l'INRIA (Institut National de la Recherche en Informatique et en Automatique) comme directeur des recherches à l'INRIA-Lorraine où il est responsable du projet SYCO (Systèmes de Compréhension et Bases de Connaissances). Il est l'auteur de plus de 250 ouvrages et articles dans le domaine de l'intelligence artificielle et de la communication homme-machine.

Anne BONNEAU
CRIN/INRIA Lorraine
B.P. 239
54506 Vandœuvre-Lès-Nancy
France

Anne Bonneau est chargée de recherches au CNRS dans la section « Sciences du Langage » au Centre de Recherche en Informatique de Nancy (CRIN-INRIA) depuis 1988. Titulaire d'un thèse de 3e cycle en Phonétique, soutenue à l'Université de Provence en 1984, puis ingénieur contractuel au Centre National d'Etudes en Télécommunications de Lannion (CNET), ses travaux portent sur l'étude des sons de la parole, en particulier la recherche d'indices acoustiques et de méthode de normalisation du locuteur, en vue d'une application à la reconnaissance automatique de la parole.

Dominique FOHR
CRIN/INRIA Lorraine
B.P. 239
54506 Vandœuvre-Lès-Nancy
France

Dominique Fohr est ingénieur de l'École Nationale Supérieure d'Électricité et de Mécanique de Nancy. Il a présenté une thèse en informatique en 1986 à Nancy sur la conception et la réalisation d'un système expert pour le décodage phonétique (APHODEX). Il est actuellement chargé de recherche au CNRS au CRIN/INRIA dans l'équipe *Reconnaissance des Formes et Intelligence Artificielle.* Son domaine de recherche concerne l'interprétation de signaux industriels et la reconnaissance de la parole.

Yves LAPRIE
CRIN/INRIA Lorraine
B.P. 239
54506 Vandœuvre-Lès-Nancy
France

Yves Laprie est enseignant-chercheur à l'Université de Nancy 1. Ingénieur civil des Mines et titulaire d'une thèse en informatique de l'INPL obtenue au Centre de Recherche en Informatique de Nancy (CRIN-INRIA) en 1990, ses activités de recherche portent sur l'approche à base de connaissances du décodage acoustico-phonétique de la parole continue et la conception de nouveaux algorithmes de suivi de formants.

(*) Ce travail a été mené avec le soutien du GRECO-PRC « Communication Homme-Machine ».

Yifan GONG
CRIN/INRIA Lorraine
B.P. 239
54506 Vandœuvre-Lès-Nancy
France

Yifan Gong est diplômé du Département de Télécommunications de l'Institut de Technologie de Nanjing. Il a obtenu son doctorat en informatique à l'Université de Nancy en 1988 dans le domaine de l'interprétation de signaux.
Y. Gong est actuellement chargé de recherches au CNRS dans l'équipe *Reconnaissance des Formes et Intelligence Artificielle* du CRIN/INRIA.
Ses principaux centres d'intérêt concernent l'interprétation automatique de signaux et la reconnaissance de la parole.

Jean-Marie PIERREL
CRIN/INRIA Lorraine
B.P. 239
54506 Vandœuvre-Lès-Nancy
France

Jean-Marie PIERREL, 38 ans, est professeur d'informatique à l'université de Nancy I. Depuis plus de 15 ans, il travaille en reconnaissance et compréhension de la parole continue au sein du CRIN (Centre de Recherche en Informatique de Nancy), URA 262 du CNRS. Après avoir successivement défini et mis en œuvre les systèmes MYRTILLE I et II, travail qui fut couronné par le prix scientifique IBM-France en informatique, il orienta ses recherches plus particulièrement vers le dialogue oral homme-machine puis aujourd'hui vers le dialogue multi-mode. Il est aujourd'hui directeur adjoint du CRIN et responsable, au sein de l'équipe *Reconnaissance des Formes et Intelligence Artificielle,* d'un projet de recherche « Parole et dialogue » commun au CRIN et à l'INRIA.

RÉSUMÉ

Le décodage acoustico-phonétique constitue une étape importante en reconnaissance de la parole continue. Cet article rappelle d'abord les difficultés du problème et les principales méthodes qui ont été proposées pour le résoudre. Il présente ensuite les diverses approches complémentaires adoptées par notre équipe : système expert fondé sur l'activité de lecture de spectrogrammes, reconnaissance par triplets phonétiques, modèle connexionniste de colonne corticale et reconnaissance par méthode stochastique sans segmentation.

MOTS CLÉS

Décodage acoustico-phonétique, reconnaissance de la parole, système expert, triplets phonétiques, modèles connexionnistes, colonnes corticales.

SUMMARY

Acoustic-phonetic decoding of speech recognition constitutes a major step in the process of continuous speech recognition. This paper reminds the difficulties of the problem together with the main methods proposed so far in order to solve it. We then concentrate on the different complementary approaches that have been investigated by our group : expert system based on spectrogram reading, recognition by phonetic triphones, connectionist model based on the cortical column unit and stochastic recognition without segmentation.

KEY WORDS

Acoustic-phonetic decoding, automatic speech recognition, expert system, phonetic triplets, neural networks, cortical column.

1. Introduction

Le processus d'interprétation d'une phrase est en général décrit comme une succession d'étapes depuis le niveau acoustique jusqu'au niveau sémantique [1], [2]. Dans ce schéma général, le niveau de décodage acoustico-phonétique (DAP) constitue une étape importante et une difficulté majeure dans la conception d'un système de reconnaissance [3]. Le DAP concerne l'ensemble des processus de transformation du signal acoustique continu en une description linguistique discrète sous forme d'unités telles que phonèmes, diphonèmes, syllabes, etc.

De nombreuses méthodes ont été proposées pour résoudre ce problème. Nous nous proposons dans cet article de rappeler les grands principes de ces méthodes et d'exposer plus en détail de façon comparative les différentes approches adoptées par notre équipe.

2. Présentation du décodage acoustico-phonétique

2.1. GÉNÉRALITÉS

Le signal acoustique de parole peut être vu comme le résultat d'un processus complexe d'encodage hiérarchique

comportant différents niveaux : pragmatique, syntaxico-sémantique, lexical, phonologique, articulatoire, etc. Ce signal possède des caractéristiques propres qui rendent difficile sa compréhension, notamment : continuité (ce qui nécessite de segmenter le signal à un moment ou un autre du processus), variabilité (qui rend d'autant plus complexe la reconnaissance multilocuteurs) et redondance.

Une langue telle que le français possède un nombre relativement faible de phonèmes mais la variété des formes acoustiques correspondantes est très grande. Cette variété est de plus difficile à caractériser du fait de divers facteurs :

— le débit de parole, l'intensité sonore et la hauteur de la voix exercent une influence sur les sons de la parole,
— les formes acoustiques dépendent par ailleurs du type de production des sons,
— les effets de coarticulation, le contexte, les particularités propres aux locuteurs produisent une énorme diversité dans les réalisations acoustiques.

Un facteur primordial en ce qui concerne le DAP est l'importance de la variabilité contextuelle des sons qui tire sa source d'un ensemble de phénomènes divers [4] : coarticulation (labiale, nasale, etc.), différences dans les longueurs du conduit vocal, variations du spectre de source, phénomènes de dévoisement, etc.

Les performances d'un système de reconnaissance automatique de la parole continue sont directement dépendantes de celles de l'étape de DAP. Il importe donc de concevoir des décodeurs phonétiques robustes et efficaces.

Le processus de DAP englobe deux tâches différentes qui peuvent être menées en séquence ou en parallèle, selon la méthode adoptée :

— une tâche de segmentation du signal vocal en unités élémentaires,
— une tâche d'étiquetage phonétique de ces segments.

Le choix d'une unité phonétique est d'une grande importance. Diverses unités ont été utilisées, parfois concurremment dans le même système : syllabes, demi-syllabes, diphonèmes, triplets, phonèmes, etc.

Syllabes, demi-syllabes, diphonèmes et triplets présentent l'avantage d'intégrer des informations sur les transitions, parties les plus difficiles à identifier dans la parole. L'utilisation d'allophones peut, dans une certaine mesure, faciliter la tâche d'étiquetage automatique mais il est nécessaire de trouver un compromis performance/complexité car le nombre d'allophones est potentiellement infini. De nombeux systèmes font intervenir le phonème comme unité, à un moment ou à un autre du décodage. Plusieurs raisons militent pour cela. Le phonème est l'unité de base de description de la parole en termes acoustiques. De plus, cette unité nécessite moins de temps d'apprentissage et moins d'espace mémoire de stockage. Enfin, les variations phonologiques des phonèmes peuvent être prédites à l'aide de règles contextuelles intra- et inter-phonèmes.

En fait, aucune unité de décodage n'est entièrement satisfaisante. Nous avons ainsi été amenés à utiliser deux types d'unités lors de nos diverses expériences : phonèmes (cf. §§ 3 et 5) et les triplets phonétiques (cf. § 4).

2.2. TECHNIQUES DE DAP

2.2.1. Segmentation

Segmenter est une opération fondamentale du DAP. Les diverses unités phonétiques déjà évoquées posent des problèmes spécifiques : *« phones »* [5], phonèmes [6], [7], diphonèmes [8], [9], demi-syllabes [10], syllabes [11], [12], [13], [14], [15] et triplets [16]. Les segments peuvent être obtenus soit de façon synchrone par analyse d'échantillons de parole de taille constante (échantillons « centi-secondes ») [17], soit de façon asynchrone [18].

Le principe de segmentation se fonde en général sur l'étude des variations d'une fonction mesurant les variations et les discontinuités de l'onde vocale et de son spectre. Des critères articulatoires ont également été utilisés [19]. La segmentation repose dans tous les cas sur une connaissance acoustico-phonétique qu'il est possible de rendre explicite. Cette démarche conduit à des systèmes de segmentation par règles de réécriture [20], [21] ou à bases de connaissances [22]. Cette dernière approche permet d'intégrer des connaissances heuristiques que possède le phonéticien et d'améliorer ainsi les performances. La plupart des segmenteurs utilisent une stratégie ascendante partant des données brutes du signal pour remonter vers une représentation plus symbolique. Une stratégie descendante fondée sur la prédiction de la succession de phonèmes a été également proposée.

Toutes ces méthodes introduisent des erreurs de sur- ou sous-segmentation qu'il importe de détecter et de corriger au mieux. Cela implique en particulier deux démarches largement répandues dans les systèmes actuels : une interaction aussi étroite que possible entre segmentation et étiquetage phonétique, une stratégie de « décision retardée » consistant à repousser des décisions définitives de segmentation (mais aussi d'étiquetage) le plus tard possible de façon à rassembler le maximum d'éléments décisifs.

Une alternative consiste à concevoir des modèles de reconnaissance sans segmentation explicite. Nous avons développé une méthode de ce type qui sera présentée au paragraphe 6.

2.2.1. Etiquetage phonétique

L'étiquetage phonétique de la parole fait appel à des techniques le plus souvent issues de la classification ou de la reconnaissance de formes. Elles ressortissent à quelques grandes catégories :

— *la quantification vectorielle :* cette technique s'appuie sur les propriétés statistiques des sons dans un certain espace de représentation. Elle est largement utilisée en codage et en synthèse de la parole ; elle présente aussi un intérêt en reconnaissance, notamment pour effectuer un premier étiquetage en grandes classes phonétiques [23]. Une telle méthode est utilisée dans le système de reconnaissance sans segmentation décrit au paragraphe 6 ;

— *la reconnaissance de formes statistiques :* un grand nombre de systèmes de DAP relèvent de méthodes classiques de reconnaissance de formes. L'identification d'un segment est menée par comparaison à un ensemble de segments de référence, constitué a priori, dans lequel les segments sont décrits par leurs paramètres acoustiques et phonétiques et les informations statistiques associées. Ces techniques impliquent la définition d'une « mesure » efficace de la « distance » entre deux formes. Les algorithmes de programmation dynamique sont souvent utilisés pour compenser les variations non linéaires de durée des segments à comparer. L'importance des effets de contextes nécessite de définir un grand nombre de formes de référence (parfois plusieurs milliers). Ceci constitue une limitation de ces méthodes, notamment en reconnaissance multilocuteurs car les formes de référence sont très dépendantes d'un locuteur ;

— *la reconnaissance de formes structurelle :* cette branche de la reconnaissance des formes est relative à la description de formes complexes par assemblage de formes primitives simples. Cette méthode a été utilisée en DAP [20] [21], mais les limitations qui viennent d'être mentionnées pour les méthodes statistiques demeurent ;

— *la modélisation stochastique :* l'opération de DAP peut être décrite formellement comme la recherche de la meilleure chaîne (ou treillis) d'unités phonétiques décrivant une phrase fournie en entrée, par comparaison de cette phrase avec toutes les concaténations possibles d'unités de référence ou de modèles de production. Ceci s'exprime aisément en termes stochastiques, notamment dans le formalisme des sources de Markov, ou plus précisément des modèles de Markov « cachés » (HMM) [17], [24]. Initialement utilisée pour la reconnaissance de mots, la méthode peut être généralisée à la reconnaissance phonétique [16], [25], [26]. Un avantage majeur en est la possibilité de rendre compte de la variabilité du signal vocal par le traitement de très gros corpus de données acoustiques, nécessaires à un bon « ajustage » du modèle markovien. Ces modèles sont actuellement parmi les plus performants pour le DAP. En revanche, s'appuyant sur une modélisation purement mathématique, ils ne permettent pas d'introduire de façon explicite des connaissances phonétiques ou phonologiques ;

— *le raisonnement fondé sur des connaissances :* l'utilisation raisonnée des connaissances disponibles concernant les divers aspects du décodage phonétique constitue une approche séduisante dont il sera question aux paragraphes 3 et 4. Cette solution a également donné lieu à plusieurs réalisations pour l'anglais [27], [28], le japonais [29] et le français [30] ;

— *les modèles connexionnistes :* ces modèles, fondés sur une modélisation plus ou moins réaliste du cortex humain, représentent une alternative intéressante, apparue récemment avec le regain d'intérêt pour les réseaux neuromimétiques. Un modèle original relevant de cette catégorie sera présenté au paragraphe 5.

3. APHODEX : un système expert de DAP

3.1. DÉMARCHE DE L'EXPERT

Dans le but d'améliorer le décodage acoustico-phonétique, nous avons entrepris d'analyser la démarche d'un phonéticien, François Lonchamp, expert en lecture de spectrogrammes de parole. Nous travaillons dans un cadre multilocuteurs, parole continue. Dans une première étape, nous avons demandé à l'expert de décoder en notre présence des spectrogrammes pour formaliser sa stratégie et les règles qu'il utilise [31].

L'expertise se décompose en fait en deux parties :

— une analyse visuelle permettant d'extraire du spectogramme les indices acoustiques pertinents,

— un raisonnement contextuel à l'aide des indices détectés pour en déduire les phonèmes prononcés.

Ces deux expertises sont intimement mêlées lors du décodage ; l'expert extrait d'abord visuellement quelques caractéristiques pertinentes puis il ébauche un raisonnement pour obtenir une première liste de phonèmes candidats. Ensuite il va chercher d'autres indices visuels pour confirmer ou infirmer ses premières hypothèses.

3.2. FORMALISATION DE L'EXPERTISE

Nous avons formalisé l'expertise sous forme de procédures d'extraction d'indices et le raisonnement de l'expert sous forme de règles de production. De plus un module de prétraitement fournit une segmentation grossière de la phrase étudiée en grandes classes.

Le module de prétraitement

Ce module détermine le début et la fin de chacun des segments phonétiques de la phrase et la classe phonétique à laquelle il appartient. Six classes ont été retenues :

— fricatives : fsvz$\int$3,
— occlusives : pbtdkg et les silences,
— noyaux vocaliques : les voyelles orales et nasales,
— occfri : les segments qui présentent des caractéristiques à la fois occlusives et fricatives (comme f par exemple),
— fricvoc : les segments qui présentent des caractéristiques à la fois vocaliques et fricatives (comme i, j et z),
— autres : les segments qui n'appartiennent à aucune des classes précitées (principalement les liquides et les nasales).

Le principe du prétraitement se fonde sur l'étude des variations temporelles de trois paramètres :

— l'énergie dans la bande de fréquences 250 Hz-3 500 Hz servant à la détermination des segments vocaliques. Chacun des pics de cette courbe vérifiant un ensemble de conditions (hauteur minimale du pic, profondeur de la vallée de part et d'autre du pic, longueur minimale du segment, etc.) est placé dans la classe des noyaux vocaliques,

— le centre de gravité du spectre pour détecter les segments fricatifs,

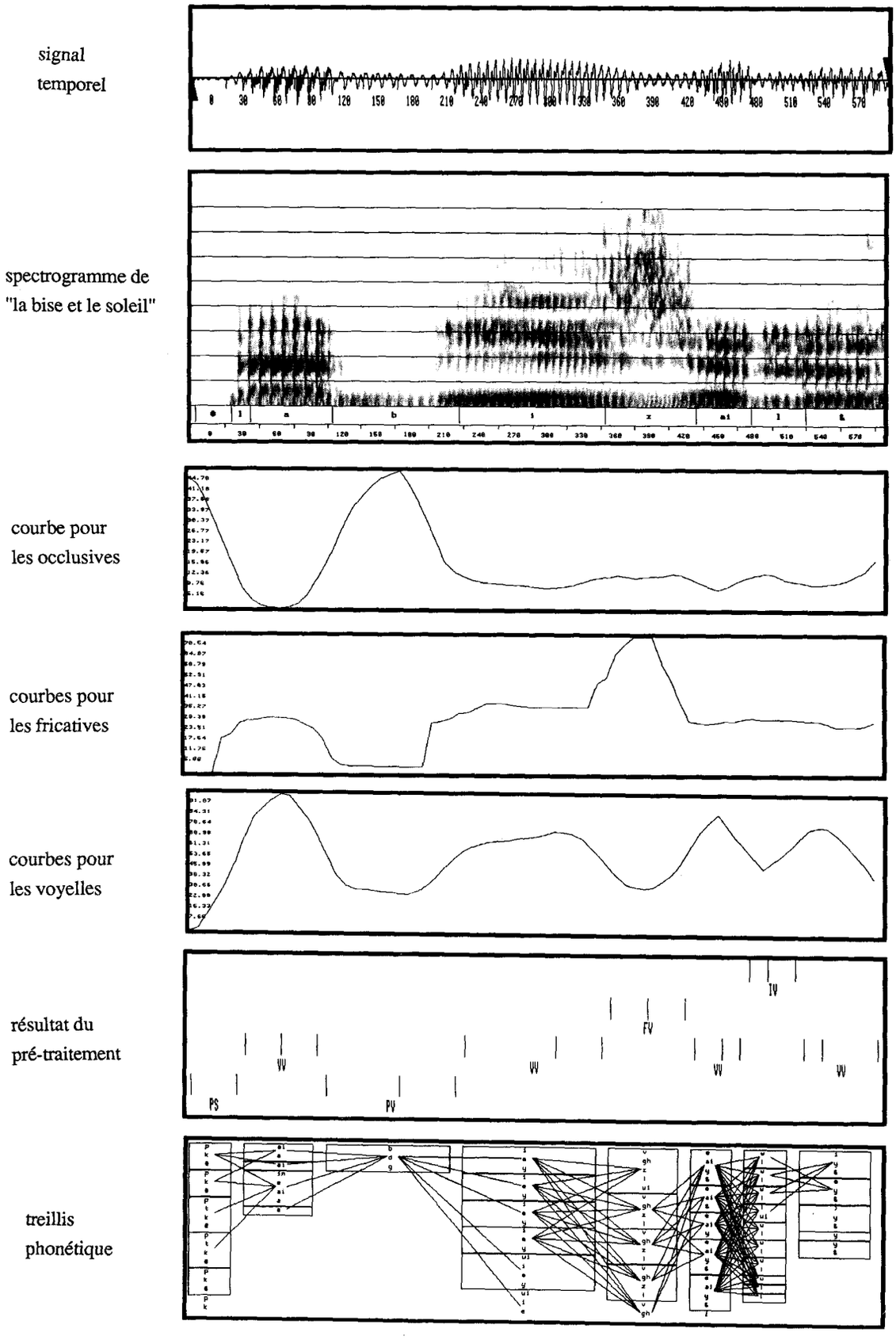


Figure 1. — Exemple de décodage pour le début de la phrase « La bise et le... » (locuteur masculin).

— l'énergie dans la bande de fréquences 500 Hz-8 000 Hz pour classer les segments de type occlusif ou silence.

Enfin, un critère de différence spectrale permet de placer les frontières entre les différents segments.

La figure 1 fournit un exemple de prétraitement obtenu à l'aide des paramètres précédents.

Les résultats de ce prétraitement sont proposés dans la figure 2. Ils ont été obtenus sur le corpus du GRECO-PRC Communication Homme-Machine « la bise et le soleil » pour dix locuteurs masculins. On peut noter un taux global supérieur à 85 % pour un taux d'insertion inférieur à 10 %, sans aucune adaptation au locuteur.

La segmentation fournie par le module de prétraitement n'est pas définitive. En effet, le moteur d'inférences peut modifier les différentes frontières trouvées et même réaliser des fusions ou des découpages de segments.

	présentes dans le corpus	trouvées	insérées
occlusives	1028	886 (86%)	21 (2%)
fricatives	521	450 (86%)	50 (10%)
voyelles	1942	1764 (91%)	77 (4%)

Figure 2. — Résultats du prétraitement.

3.3. LE SYSTÈME EXPERT [32]

Une des caractéristiques essentielles de l'expertise du phonéticien réside dans l'importance cruciale de la notion de contexte. En effet, pratiquement toutes les règles que l'expert utilise sont contextuelles. Nous avons donc développé notre propre moteur d'inférences pour tenir compte de cet impératif [33].

Les principales caractéristiques du système sont les suivantes :

— remettre en cause la segmentation à tout moment,
— dérouler en parallèle une analyse sur plusieurs segmentations possibles,
— prendre en compte les phénomènes contextuels,
— tenir compte de l'incertitude en ce qui concerne l'interprétation des mesures (détection d'indices),
— pouvoir facilement ajouter de nouvelles règles, avec interface conversationnelle pour les entrer,
— utiliser une base de connaissances compréhensible et modifiable par l'expert,
— fournir une trace du raisonnement.

Les règles

La connaissance de l'expert est formalisée sous forme de règles de production. Il existe deux sortes de règles : les règles action et les règles déduction. Les premières déclenchent une fonction qui va modifier le treillis phonétique ou la segmentation. Les autres contiennent en partie « conclusion » une liste de phonèmes pondérés par des coefficients de vraisemblance (CV).

Voici un exemple de règle :

```
R24                                 Règle 24
/*règle de la pince vélaire*/
CONTEXTE_GAUCHE [a]                 si le contexte gauche est /a/
SI
  Formant2_montant_Pre &            et si le formant 2 de ce /a/ est descendant
  Formant3_descendant_Pre           &et si le formant 3 de ce /a/ est montant
  burst_concentre_Act               et si le burst est concentré
ALORS                               ALORS
  PHONEMES [k 9 g 9]                c'est sûrement un /k/ ou un /g/ (avec un
                                    CV de 9 ; les CV varient de – 10,
                                    totalement faux à + 10, totalement vrai)
```

Comme on l'a vu, les règles sont pour la plupart contextuelles. Les listes de phonèmes décrivant les deux contextes (voisins immédiats) à droite et à gauche du phonème sur lequel porte la règle font partie des conditions d'application de la règle.

Les prémisses contiennent les différentes conditions que doivent remplir les indices détectés sur le segment. Ils peuvent être soit booléens (la présence ou l'absence de burst dans un segment par exemple) soit numériques (des fréquences de formants par exemple). On utilise une logique floue pour pallier les incertitudes liées aux seuils.

3.4. LE TREILLIS

Le résultat du décodage est un véritable treillis de phonèmes ; chaque nœud de ce treillis contient la liste ordonnée des phonèmes trouvés. Les règles ayant conduit à ces déductions étant contextuelles, on conserve pour chaque nœud les contextes gauches et droits imposés par ces règles. On peut ensuite visualiser les chemins possibles (ceux pour lesquels les contraintes dues au contexte sont satisfaites). C'est aux modules de niveau supérieur (syntaxe, sémantique, lexique...) de choisir parmi les chemins proposés celui qui correspond à une phrase possible.

Un exemple de résultat est donné en figure 1. On y trouve le signal temporel, le spectrogramme, l'étiquetage manuel de l'expert, la segmentation du module de prétraitement et le treillis phonétique résultat.

De plus, le moteur d'inférences peut fonctionner en vérification d'hypothèses. Si les niveaux supérieurs émettent l'hypothèse d'une omission d'un segment, ils peuvent indiquer à APHODEX l'endroit de l'omission, le phonème supposé omis et son contexte. Le système expert va alors créer un nouveau nœud dans le treillis avec le contexte donné par les niveaux supérieurs et appliquer en chaînage arrière les règles qui confirment ou infirment la présence du phonème émis comme hypothèse dans ce contexte particulier.

La matrice de confusion-insertion-omission obtenue sur un corpus de parole continue pour 4 locuteurs masculins est donnée figure 3. Ici encore, aucune adaptation au locuteur n'est effectuée. On peut remarquer les bons résultats sur les phonèmes pbʃszʒ ainsi que sur les voyelles. Les scores, nettement moins bons, sur les liquides et sur les nasales s'expliquent par le fait que la base de connaissances est encore incomplète pour ces

	nb	p	b	t	d	k	g	f	v	ch	gh	s	z	m	n	nj	j	w	R	l	ui	on	an	in	un	i	e	ai	a	)	o	u	y	eu	oe	&	#	omis		
p	542	441	.	.	81	.	2	.	.	.	.	.	.	.	.	.	.	1	.	1	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	1	14	81%	p
b	514	.	417	28	.	1	.	.	.	.	.	.	.	4	1	.	1	18	.	.	13	.	.	.	.	.	.	.	.	.	3	1	.	.	.	.	27	81%	b
t	966	4	223	651	.	.	7	.	.	2	.	1	.	.	.	.	1	3	2	3	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	67	67%	t
d	1079	116	58	1	823	1	.	.	.	1	1	.	.	6	7	.	7	55	1	3	6	.	.	1	.	2	1	.	.	1	.	1	2	.	.	.	2	183	58%	d
k	608	5	95	2	12	421	.	.	1	.	1	.	.	1	.	.	1	5	2	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	60	69%	k
g	422	43	26	10	.	.	196	.	.	.	.	.	.	.	.	.	6	31	1	3	4	.	.	.	.	4	1	.	.	.	.	.	.	.	.	.	97	46%	g
f	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0		f
v	185	4	30	.	.	.	.	.	12	1	.	.	.	5	.	.	29	17	57	4	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	22	6%	v
ch	64	.	.	.	.	.	.	.	.	61	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	95%	ch
gh	82	.	.	.	.	.	.	2	.	13	63	.	.	.	.	.	.	2	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	1	77%	gh
s	352	.	.	.	.	.	.	.	2	.	18	305	21	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	5	87%	s
z	188	1	.	.	.	.	.	4	.	.	.	32	128	.	1	.	1	2	2	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	1	6	71%	z
m	332	4	16	1	.	.	.	.	.	.	.	.	.	153	18	.	6	55	2	.	1	1	.	3	.	6	1	.	1	.	.	3	1	.	.	.	60	46%	m
n	401	.	35	.	3	.	.	.	.	.	.	.	.	41	67	.	25	123	1	6	1	.	.	3	.	1	.	.	.	1	2	.	.	.	.	2	89	17%	n
nj	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0		nj
j	137	.	.	.	.	.	.	.	2	.	5	.	.	.	.	.	5	22	.	10	17	.	.	.	.	4	5	.	.	.	.	.	.	.	.	.	67	4%	j
w	137	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	92	2	.	1	.	2	.	.	.	.	.	.	1	1	1	.	.	.	.	36	67%	w
R	1534	28	38	4	.	1	.	51	10	134	26	88	6	3	.	.	8	109	560	23	7	2	3	10	.	4	15	5	4	4	8	1	1	.	.	37	330	37%	R
l	1063	15	28	3	1	.	.	.	1	4	2	1	.	1	5	.	11	13	7	501	1	.	.	.	.	3	.	.	.	.	2	.	1	.	.	38	425	47%	l
ui	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0		ui
on	145	.	1	.	1	.	.	.	.	.	.	.	.	1	.	.	2	.	.	.	.	84	2	24	.	7	7	4	.	.	3	2	.	.	.	.	6	58%	on
an	369	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	4	.	.	8	261	42	3	1	13	2	.	1	8	2	2	.	.	2	.	18	71%	an
in	184	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	4	.	.	.	.	157	1	3	5	5	.	.	.	.	.	.	.	.	1	5	85%	in
un	149	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	.	.	.	8	.	130	1	1	1	.	1	.	.	.	.	.	.	.	5	87%	un
i	840	3	3	.	2	.	.	1	.	4	17	1	4	.	.	.	7	2	.	1	35	.	.	.	.	688	22	1	.	.	.	1	2	.	.	.	1	43	82%	i
e	365	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	354	.	.	.	.	.	.	.	.	.	.	7	97%	e
ai	693	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	1	1	5	.	.	.	.	.	.	.	6	649	18	2	.	.	.	.	.	1	.	16	94%	ai
a	1503	1	1	.	.	.	.	.	.	.	1	.	.	.	.	.	1	1	17	.	1	.	.	1	1	1	1	621	358	8	.	.	.	.	.	7	.	36	98%	a
)	449	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	2	.	2	.	.	5	.	28	3	389	5	.	.	.	.	8	.	4	87%	)
o	217	1	1	.	.	.	.	.	.	.	.	.	.	.	2	.	8	24	5	.	5	.	.	.	.	6	3	.	.	5	138	.	1	.	.	.	2	24	88%	o
u	618	2	12	1	.	.	.	.	.	.	1	.	.	5	5	.	6	5	5	.	6	.	.	3	1	43	3	1	.	2	12	383	28	.	3	4	1	69	64%	u
y	247	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	3	.	.	1	.	12	7	.	.	.	.	.	218	.	.	.	.	5	88%	y
eu	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0		eu
oe	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0		oe
&	1252	8	4	1	.	.	.	.	.	.	.	.	.	.	1	.	6	2	26	3	.	.	1	4	.	6	3	23	15	2	.	.	1	.	11051	2	88	84%	&
#	1230	136	3	28	1	1	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	1	.	.	.	898	59	81%	#
ins		47	22	5	5	0	0	9	14	12	22	13	4	14	11	0	50	210	133	10	67	25	15	40	1	71	40	13	19	40	36	25	16	0	0	10	10		69	%

Figure 3. — Matrice de confusion-insertion-omission.

classes de phonèmes. Il faut encore améliorer la qualité des procédures d'extraction d'indices acoustiques et le nombre de règles pour obtenir un score acceptable pour l'ensemble des classes de phonèmes.

3.5. NORMALISATION FORMANTIQUE DES LOCUTEURS

Pour améliorer les performances de reconnaissance, nous avons élaboré, d'après une idée de F. Lonchamp, une méthode de normalisation des locuteurs fondée sur un apprentissage minimal grâce à l'apport de connaissances. A partir des valeurs formantiques d'une voyelle donnée qui constituent les données d'apprentissage, ainsi que des valeurs formantiques moyennes des hommes et des femmes, nous déterminons des paramètres de normalisation qui tiennent compte implicitement de la longueur relative du conduit vocal de chaque locuteur.

Les écarts fréquentiels entre les formants des hommes et des femmes sont principalement dus à la longueur du conduit vocal ; supérieure de 17 % environ chez les hommes, elle entraîne une chute des fréquences masculines presque équivalente.

Rappelons que ceci est valable en moyenne et s'applique relativement bien aux voyelles centrales. Pour les autres voyelles, les écarts entre formants ne sont pas homogènes quels que soient la voyelle et le formant considérés, ainsi que Fant [34] l'a souligné. Le conduit vocal de l'homme est en effet plus allongé dans le pharynx (et les cavités laryngées) que dans la partie buccale ; ceci explique la chute plus importante des fréquences des formants qui sont particulièrement dépendants du pharynx — par exemple, les deuxièmes formants des voyelles antérieures. D'autres facteurs influent, notamment l'ouverture au point de constriction maximale [35], mais sont ici ignorés, des phénomènes de compensation rendant leur détection délicate.

Afin de respecter les conséquences de ces différences de longueur sur chaque formant, notre méthode est fondée sur des connaissances : pour chaque voyelle, nous disposons en effet des fréquences moyennes des trois premiers formants pour les hommes et pour les femmes, calculées sur des mots isolés. Prenons ces valeurs comme références et appelons-les rh_j et rf_j pour représenter respectivement les moyennes masculines et féminines de la voyelle j (*fig.* 4).

Nous faisons l'hypothèse suivante : une évaluation, pour une voyelle quelconque, des valeurs formantiques d'un locuteur par rapport à celles de nos références, nous renseigne sur la longueur relative du conduit vocal du locuteur par rapport à la longueur des conduits moyens des hommes et des femmes. Grâce, d'une part, à cette information, indépendante de la voyelle choisie et, d'autre part, aux valeurs de référence rh_j et rf_j de chaque voyelle j du français, nous pouvons déduire les valeurs formantiques typiques du locuteur pour chaque voyelle. Ces valeurs deviendront les références caractéristiques du locuteur considéré.

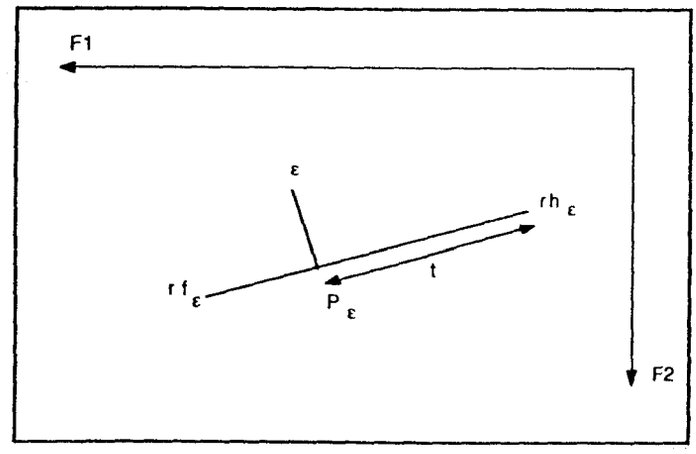


Figure 4. — Projection de la voyelle /ε/ dans le plan F1-F2 (les calculs ont été effectués en réalité dans l'espace F1-F2-F3), et détermination du paramètre de normalisation t, représenté ici dans le cas où la norme du vecteur $rh_\varepsilon rf_\varepsilon$ vaut 1.
rh_ε et rf_ε : références masculines et féminines.
P_ε : projection de ε.

Soit t la longueur relative estimée pour le locuteur l, et rl_j la référence à déterminer pour ce locuteur l et la voyelle j, $\overrightarrow{rh_j rl_j} = t * \overrightarrow{rh_j rf_j}$.

Nous avons testé notre procédure sur les voyelles orales du français, en parole continue, pour des locuteurs des deux sexes.

Les valeurs moyennes des voyelles ont été extraites d'un corpus de mots isolés, d'où un léger problème d'estimation de nos paramètres de normalisation. Malgré cela, nos nouvelles références individualisées entraînent une réduction de variance d'environ 50 %, et donc une amélioration de la reconnaissance.

4. Le triplet phonétique, une unité contextuelle

4.1. DÉFINITION ET INTÉRÊT DU TRIPLET

L'unité de décodage d'APHODEX est le phone, et l'expertise modélise la réalisation acoustique du phone en prenant en compte son contexte phonétique. Le fait que le grain de connaissance nécessaire au décodage d'un phone fasse intervenir des indices contextuels et donc dépendant des phones voisins, engendre un certain nombre de problèmes liés à la cohérence de l'expertise. Il faut ainsi assurer qu'une nouvelle règle ne risque pas de contredire, même partiellement, celles déjà présentes dans la base de connaissances.

Comme l'expertise acoustico-phonétique est destinée à la classification de phones, il est naturel de construire des prototypes représentant les phones à reconnaître. Pour que ces prototypes soient utilisables, ils doivent modéliser la réalisation acoustique en prenant en compte les phénomènes de coarticulation liés à la présence des phones voisins. Cela nous a donc conduit à proposer un système

dont le grain de connaissance est le triplet : un phone avec son contexte phonétique.

Il est important de noter qu'un triplet ne doit pas inclure les phonèmes voisins en entier sinon il devient à son tour soumis aux phénomènes contextuels auxquels on veut précisément échapper.

Un triplet s'articule donc autour des deux frontières du phone central comme le montre la figure 5. A chacune des frontières sont attachés les événements acoustiques suivants s'ils sont présents :

— transitions formantiques,
— burst avec sa durée, son intensité et la localisation fréquentielle des concentrations d'énergie,
— limite inférieure de bruit (pour les fricatives),
— points d'échange des cavités formantiques,
— profil de la micromélodie.

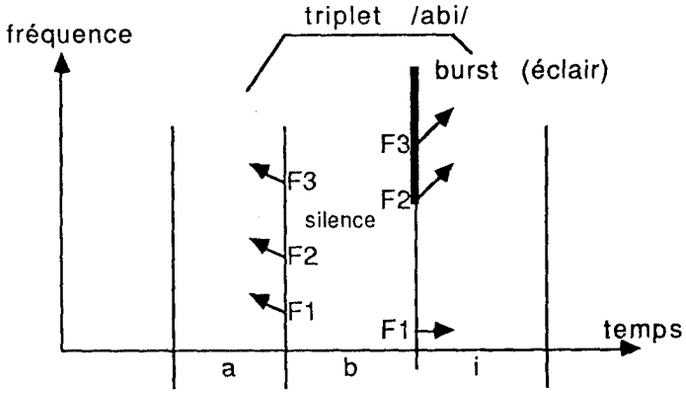


Figure 5. — Exemple de description d'un triplet.

Pour que les références fréquentielles soient suffisamment précises nous avons ajouté un « centre de triplet » qui indique les fréquences des formants au centre du triplet.

Tel que nous venons de le décrire, un triplet n'est que la description acoustique d'un phonème en contexte et ne porte donc aucune information sur l'expertise qu'un lecteur de spectrogramme a pu accumuler. La description acoustique du triplet est donc complétée par les indices qu'utilise un expert phonéticien. Un indice est un corrélat acoustique indépendant du locuteur, résistant et reconnu par l'expert comme significatif ; c'est par exemple :

— une pince vélaire (rapprochement de F2 et F3 à la frontière d'une occlusive vélaire),
— la position relative de la concentration d'énergie d'un burst par rapport à celle des formants.

Les deux facettes de cette représentation permettent d'orienter le décodage soit vers la recherche des triplets en fonction d'indices, quand les indices apparaissent clairement dans le signal, soit vers la recherche des triplets en fonction de leur réalisation acoustique quand peu d'indices décisifs ont été découverts.

La facette faisant appel aux indices d'un triplet n'est d'ailleurs pas toujours présente, soit qu'elle n'ait pas été construite par l'expert, soit que l'état des connaissances acoustico-phonétiques ne permette pas de définir des indices pertinents. Cette facette est cependant très intéressante car elle permet de reconnaître avec plus de certitude les triplets qui contiennent des indices très caractéristiques, ceux que l'on commence maintenant à bien connaître.

A l'intérêt que représente le triplet pour exprimer les connaissances acoustico-phonétiques et assurer la cohérence de l'expertise, s'ajoute la possibilité d'adapter le décodage en fonction du locuteur. Il est évidemment possible de normaliser, notamment en fréquence, les triplets de la base de connaissances mais cette adaptation ne peut être que grossière puisque l'on veut qu'elle soit très rapide. Mais on peut en fait aller beaucoup plus loin. Il est en effet possible de vérifier que les relations fréquentielles existant entre deux triplets de la base de connaissances, proposés comme solutions du décodage pour deux segments à décoder, existent aussi entre leurs instances dans la phrase à décoder. Nous reviendrons en détail sur ce point au paragraphe 4.4.2 qui permet de tirer profit de l'unicité du locuteur pour construire un décodage global et cohérent.

4.2. ORGANISATION DES CONNAISSANCES

Le volume de connaissances à acquérir est évidemment considérable même si de nombreux triplets sont impossibles. L'étude de Tubach et Boé [36] permet de constater qu'environ 1 750 triplets suffisent à représenter 75 % des triplets d'un très vaste corpus (300 000 phonèmes). Il reste que pour valider l'intérêt du triplet en décodage, nous nous limiterons à un sous-ensemble de l'ordre de 500 triplets représentant essentiellement des occlusives et des sonantes, phonèmes assez difficiles à reconnaître avec APHODEX. Notons que ce nombre de triplets ne représente que la moitié des triplets construits à partir des occlusives en contexte vocalique.

Les triplets sont organisés suivant leur profil phonétique (le triplet de la figure 5 appartient donc à la classe [Voyelle Occlusive voisée Voyelle]). Cela représente encore un grand nombre de triplets par classe ([Voyelle Occlusive voisée Voyelle] conduit à $13 * 3 * 13 = 507$ triplets). Nous verrons au paragraphe 4.4.1 comment affiner cette organisation.

La réduction du nombre des triplets (et donc de l'effort d'acquisition des connaissances acoustico-phonétiques), destinée à faciliter l'évaluation de cette approche, ne doit pas occulter le fait qu'une très vaste connaissance des phénomènes de parole doit être disponible pour améliorer les performances de l'approche experte, cela quelle que soit la manière dont l'expertise est formalisée.

4.3. ACQUISITION DES CONNAISSANCES

L'acquisition des connaissances est scindée en deux étapes. La première consiste à collecter les triplets d'un corpus. La description est effectuée grâce à un éditeur de triplets convivial que nous avons développé sur le noyau du système SNORRI [37]. La description a lieu directement sur le spectrogramme, l'utilisateur « dessinant » avec la

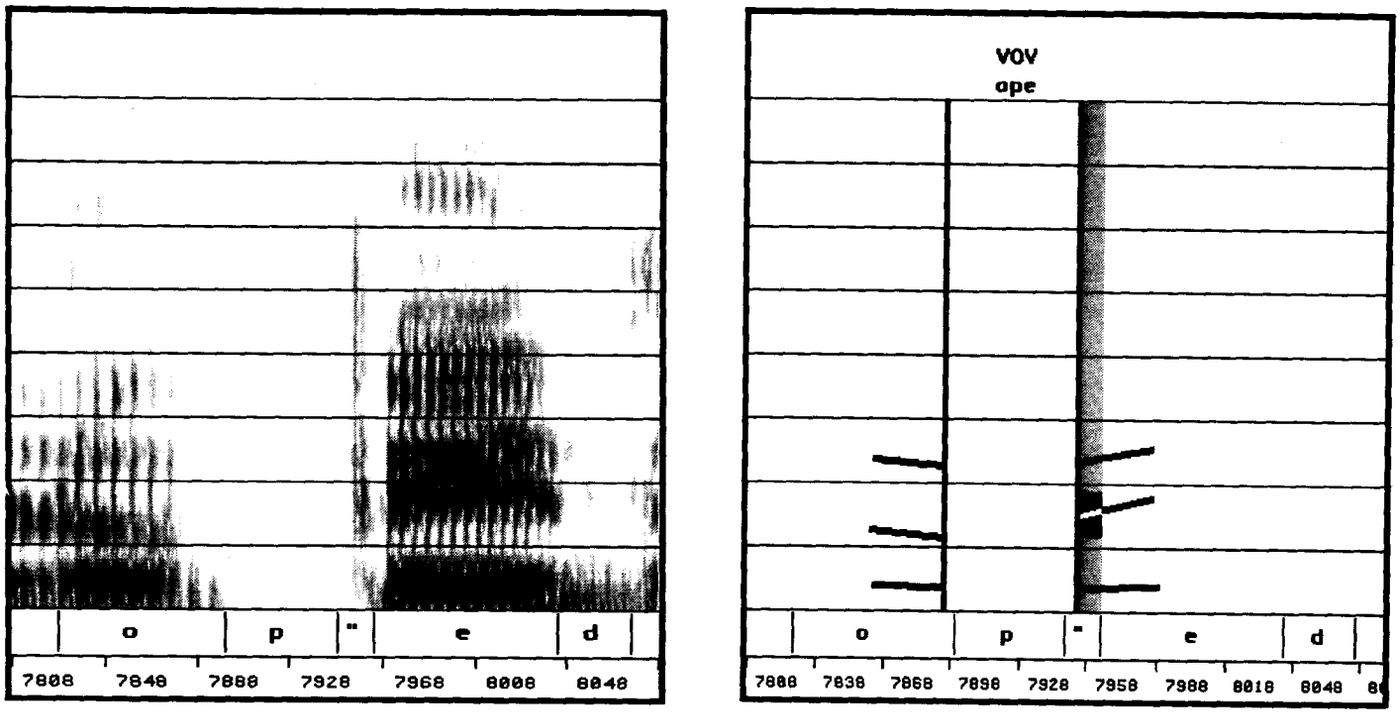


Figure 6. — Exemple de construction d'un triplet à partir du spectrogramme (frontières, formants, barre d'explosion et concentration d'énergie de la barre d'explosion, profil phonétique).

souris les différents événements acoustiques. La figure 6 illustre ce processus. Il est possible de procéder de manière semi-automatique pour les formants, en utilisant le suivi que construit SNORRI et en le corrigeant éventuellement ensuite. Comme il est essentiel que l'apprentissage se fasse sur des données correctes, nous préférons, pour l'instant du moins, conserver une partie manuelle qui permet de s'assurer de la pertinence des triplets collectés.

Comme SNORRI permet d'extraire une séquence de phonèmes donnée d'un corpus étiqueté, l'utilisateur peut collecter tous les triplets ayant le même profil phonétique et donc concentrer son attention sur une classe de triplets bien précise.

La seconde partie de l'apprentissage consiste à élaborer les triplets de référence à partir des triplets qui viennent d'être collectés. C'est là qu'intervient l'expertise du phonéticien qui permet de retenir les événements acoustiques pertinents et de définir parallèlement les corrélats acoustiques qui serviront comme indices de décodage.

Les corpus dont nous disposons (« La bise et le soleil » extrait du corpus BDSONS, ainsi que le corpus d'O. Mella [38]) et grâce auxquels nous construisons les triplets de test ne contiennent qu'environ un millier de triplets. Pour étendre cette base de connaissances il faudra donc, soit disposer de nombreux autres corpus, soit pouvoir construire directement les triplets de référence.

Comme l'apprentissage, en dehors de celui que nous allons réaliser pour le test, représente un volume de travail gigantesque, nous étudions la possibilité de le réaliser automatiquement. La première partie de l'apprentissage est tout à fait automatisable en remplaçant le phonéticien ou le chercheur en parole par les détecteurs automatiques d'indices dont nous disposons déjà pour le système APHODEX. La seconde partie, consistant à regrouper les descriptions disponibles pour en extraire des modèles représentatifs pourrait être réalisée en utilisant des techniques de classification conceptuelle [39].

4.4. PROBLÈMES DE RECONNAISSANCE

Le processus de décodage qui va être maintenant décrit intervient après la segmentation et la détection des événements acoustiques. Cette étape préliminaire est destinée à construire à partir de la phrase une suite d'instances de triplets qui seront identifiés grâce aux triplets de référence.

Le décodage proprement dit se décompose en deux étapes successives. La première conduit à proposer, pour chaque instance de triplet, la liste des triplets de référence qui s'apparient le mieux avec le segment inconnu. La seconde étape est destinée à augmenter la cohérence de la solution globale pour la phrase en assurant que les contraintes qui existent entre les triplets de la base de connaissances qui ont été proposés comme solutions sont aussi vérifiées par les instances de triplets de la phrase à décoder.

4.4.1. Décodage grossier

Ce décodage est essentiellement un décodage local et repose sur l'appariement du triplet inconnu aux triplets de la base de connaissances qui en sont les plus proches. Précisons néanmoins que, connaissant la classe phonétique

du triplet, seuls les triplets de même profil phonétique sont concernés par cet appariement. Pour que la comparaison soit significative, il est nécessaire de prendre en compte la variabilité interlocuteurs et donc d'effectuer une normalisation des triplets par rapport au locuteur qui a énoncé la phrase à décoder. Cette normalisation doit être rapide (tout au plus quelques mots) : nous avons choisi la méthode de modélisation implicite de la longueur du conduit vocal [40] décrite au paragraphe 3.5. Nous limitons pour l'instant la normalisation aux fréquences, mais il serait possible de normaliser les triplets en durée en modélisant les phénomènes de réduction liés à la vitesse d'élocution [41].

L'un des points clés d'un système à base de prototypes [42] est l'organisation des connaissances afin que la recherche des triplets les plus proches du triplet inconnu ne nécessite de parcourir qu'une petite partie de la base de connaissances. La solution couramment adoptée en image [43], comme en parole [44] est de hiérarchiser les prototypes suivant leurs indices caractéristiques. S'il est possible de structurer de cette manière une partie des triplets — c'est-à-dire ceux pour lesquels nos connaissances actuelles des phénomènes de parole permettent de définir des indices acoustiques clairs — il reste que pour une partie importante des triplets peu d'indices décisifs en reconnaissance sont connus. Il est cependant très important de structurer de cette manière une partie des connaissances, car l'expert ne place pas toutes ses connaissances au même niveau. Prenons l'exemple du triplet /syR/ ; la forte descente de F2 vers F1 qui est un indice très caractéristique permet à l'expert de conclure rapidement qu'il s'agit d'un triplet se terminant par /R/ et centré sur une voyelle proche de /y/. Quand ce « raccourci » heuristique de raisonnement n'est pas possible, ce qui est fréquemment le cas, il faut recourir au second mode d'organisation des connaissances qui repose cette fois sur les profils formantiques (et donc sur les références fréquentielles des triplets). Les triplets sont classés suivant les fréquences des formants F1, F2, F3 aux frontières du phone central. Le classement est effectué par classe d'amplitude relativement large pour tenir compte de l'imprécision des fréquences due aux fortes transitions formantiques. Le mode de structuration donne lieu à un mode de raisonnement tout à fait différent du précédent : les triplets sont cette fois comparés grâce à leur réalisation acoustique. Les indices acoustiques, définis par le phonéticien, ne sont utilisés dans ce cas que pour moduler le résultat de la comparaison.

Ces deux modes de raisonnement, l'un guidé par les indices acoustiques, l'autre guidé par les réalisations acoustiques ne sont, tels que nous venons de les décrire, utilisables que si une segmentation de la phrase est disponible. Dans le cas où aucune segmentation préexiste au décodage, ou qu'il est indispensable de la remettre en cause, la structuration suivant les indices permet néanmoins de construire une solution. Les indices sont d'abord systématiquement recherchés dans la partie à décoder ; les indices découverts permettent alors de segmenter la parole et de proposer une solution composée de triplets caractérisés par ces indices. Cette démarche est comparable à celle qu'utilisent P. D. Green *et al.* dont le système tente de reconstruire des « sketchs acoustiques » à partir des trajectoires formantiques [45].

Le résultat final du décodage grossier est donc pour chaque segment la liste des triplets les plus proches, soit en termes d'indices acoustiques, soit en termes de réalisations acoustiques.

4.4.2. Cohérence globale de la solution

Le résultat du décodage précédent n'est que la juxtaposition des solutions locales pour chacun des segments de la phrase. Il est donc possible que les contraintes liant deux triplets de références (une contrainte portant sur les positions relatives des formants des deux triplets) ne se retrouvent pas sur les instances des triplets de la phrase à décoder. Éliminer de telles incohérences, permet donc de proposer une solution consistante sur toute la phrase et donc en somme de modéliser l'unicité du locuteur dans le processus de décodage.

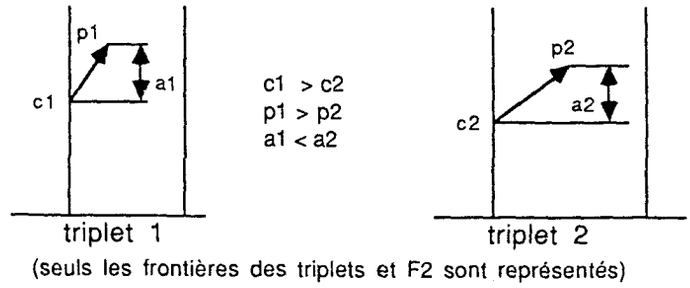
Le problème à résoudre est le suivant. Étant donnés n nœuds (les segments de parole) notés i et les ensembles d'étiquettes Li proposés pour chaque nœud (l'ensemble des étiquettes étant l'ensemble des triplets proposés pour le segment i) comment éliminer les étiquettes incompatibles avec celles des autres nœuds de la phrase ? Il s'agit là du schéma classique de la relaxation discrète, connue aussi sous le nom de filtrage de Waltz [46]. Il existe deux types de consistance que l'on peut atteindre :

— consistance globale : les contraintes sont satisfaites simultanément pour l'ensemble des étiquettes de tous les nœuds ;
— consistance sur les arcs : cette consistance est plus faible que la précédente et assure que les étiquettes de deux nœuds contraints l'un par l'autre sont compatibles entre elles.

C'est le second type de consistance qui nous intéresse et de nombreux algorithmes existent pour résoudre ce problème.

Les contraintes sont construites sur les profils formantiques et font intervenir trois critères (*fig.* 7) :

— positions relatives des trajectoires formantiques aux frontières du triplet ($c1 > c2$),
— relation entre les pentes des transitions formantiques ($p1 > p2$),
— relation entre les amplitudes fréquentielles des transitions ($a2 > a1$).



(seuls les frontières des triplets et F2 sont représentés)

Figure 7. — Sous-contrainte liant les formants F2 de deux triplets.

Le fait que certains segments de parole soient mal placés ou bruités risque de conduire à un étiquetage global vide et il est donc nécessaire de disposer d'un algorithme de relaxation flou. Nous avons choisi GaC4 (développé dans notre équipe pour la vision par ordinateur) et sa version floue [47] qui permet d'éviter le risque d'éliminer toutes les étiquettes.

Cette étape de relaxation, après le décodage grossier présente deux avantages importants :

— elle permet d'approcher le comportement d'un expert qui propose une solution cohérente pour toute la phrase,

— grâce aux relations de fréquences qu'imposent les contraintes, il n'est pas nécessaire pour le décodage grossier d'effectuer une comparaison fréquentielle très précise.

Le choix du triplet comme unité de décodage repose essentiellement sur sa bonne résistance aux effets contextuels. Nous estimons que les variations dues notamment à la position des triplets par rapport aux frontières de mots sont suffisamment faibles pour être négligées, du moins en première approche. Les modifications de durée et celles dues au phonème de réduction seraient sans doute assez faciles à prendre en compte avec un modèle suffisamment éprouvé.

De nombreuses applications des systèmes de reconnaissance de parole portent sur un vocabulaire limité ; il est donc envisageable de construire un système de décodage acoustico-phonétique n'utilisant que le triplet. L'étape d'apprentissage même si elle reste importante est alors techniquement réalisable ; ainsi Lee pour un système de dialogue limité (1 000 mots) a recensé 2 380 triplets [48]. Une autre solution est d'utiliser un système à base de triplets en coopération avec APHODEX pour décoder une classe phonétique particulière.

La version du système réalisant le décodage grossier utilise, à l'heure actuelle, de manière simpliste la segmentation du système APHODEX car nous n'abordons dans un premier temps que des phonèmes par lesquels la segmentation est relativement simple (les occlusives). Il faudra cependant compléter le système de décodage par des stratégies permettant de remettre en cause une segmentation qui semble erronée.

5. Le modèle connexionniste de colonne corticale

5.1. INTRODUCTION

Des résultats prometteurs ont récemment suscité un regain d'intérêt pour l'approche connexionniste en reconnaissance automatique de la parole : on voit ainsi des applications en prétraitement du signal [49], en détection de traits phonétiques grossiers [50] et en reconnaissance de phonèmes ou de mots [51]. Les réseaux neuro-mimétiques possèdent des qualités intéressantes : parallélisme, robustesse vis-à-vis de données bruitées ou incomplètes, pouvoir discriminant, capacité de généralisation sur peu d'exemples appris, même s'ils n'ont pas encore démontré leur supériorité par rapport aux méthodes classiques.

Une difficulté du décodage acoustico-phonétique réside dans le caractère temporel de la parole (influence contextuelle, continuité des données, variation de la vitesse d'élocution, etc.), mieux maîtrisé par les techniques classiques que par les réseaux de neurones formels (ainsi les perceptrons ont été conçus pour des données statiques). C'est pourquoi le connexionnisme s'oriente actuellement vers la prise en compte de cette dimension temporelle [52], [53], [54].

Nous allons voir comment le modèle original que nous proposons aborde ces problèmes, après un bref rappel des caractéristiques essentielles de ce modèle [55], [56].

5.2. CARACTÉRISTIQUES DU MODÈLE

Nous avons adopté le modèle biologique de la colonne corticale élaboré par Y. Burnod [57], après l'avoir adapté pour une simulation informatique. Ce modèle a déjà été décrit par ailleurs [56], [58] ; en voici les points essentiels :

— l'unité élémentaire, colonne corticale, possède trois couches pour traiter stimuli, actions et concepts. Elle correspond à une fonctionnalité de plus haut niveau que les unités « neuronales » que l'on rencontre dans les perceptrons, par exemple ;

— trois états d'activité représentent inhibition (E0), déclenchement (E2) et recherche active (E1) ;

— les règles de fonctionnement sont régies par un tableau d'entrée/sortie [58] ;

— l'apprentissage est un reflet du couplage ou du découplage entre colonnes ; il peut développer trois types de relation entre deux colonnes (cf. *fig.* 8 et 9) : inhibition, déclenchement conditionnel (gating), déclenchement inconditionnel (triggering).

L'apprentissage est mis en œuvre sélectivement ; la plausibilité biologique du modèle le rend particulièrement bien adapté à traiter des problèmes typiquement humains tels que la perception de la parole ou l'interprétation d'images.

Un réseau de reconnaissance sensorielle nécessite le plus souvent trois aires par analogie avec l'organisation du cortex humain : une aire perceptive recevant et mettant en forme les signaux d'entrée, une aire de sortie et une aire associative chargée d'associer une configuration d'entrée à un réponse, comme on le verra dans l'application décrite au paragraphe 5.4.

5.3. LA DIFFÉRENCIATION DANS L'AIRE ASSOCIATIVE

5.3.1. Exposé du problème

Un expert phonéticien repère sur un spectrogramme des domaines spécifiques de fréquence lui permettant de différencier les phonèmes par la présence ou l'absence de signal et son intensité. La solution consiste donc pour effectuer simplement une reconnaissance automatique, à relier un phonème donné de la couche de sortie au domaine de fréquence qu'inspecte le phonéticien pour

prendre sa décision. Le premier moyen qui vient à l'esprit est d'effectuer cette opération « à la main », en demandant à l'expert, pour chaque phonème, ce qu'il regarde. A cette technique, on peut objecter plusieurs remarques : outre son caractère fastidieux, cette entreprise n'est pas optimale. Un expert peut ne pas reconnaître tous les points de différenciation, ou ne pas en être conscient. De plus, il ne pourra définir précisément les limites fréquentielles d'influence d'un phonème sur son contexte. Enfin, l'objection principale : cette connectivité ne serait applicable qu'à ce problème et devrait être refaite pour une autre application. D'où l'idée de réaliser une différenciation automatique, par apprentissage.

5.3.2. Données neurobiologiques

Différentes études ont démontré la plasticité des cartes dans le cortex. Cette plasticité est définie comme une corrélation entre l'activité due aux stimuli sensoriels et celle d'une autre activation corticale. Burnod [57] a formalisé la réorganisation fonctionnelle de la surface corticale de la manière suivante : une aire associative, recombinaison d'une aire sensorielle et d'une aire motrice, n'est pas *a priori* différenciée. Quand l'activité des aires sensorimotrices active un module de l'aire associative à l'état E1, ce module appelle une action de différenciation qui, tout en partageant la même surface, crée deux modules avec une activité opposée (E2-E0). Chaque nouveau module peut à nouveau être différencié selon ce même principe. Ce phénomène, illustré par la figure 8, fait tendre la carte qui le supporte vers un état de contraste maximal.

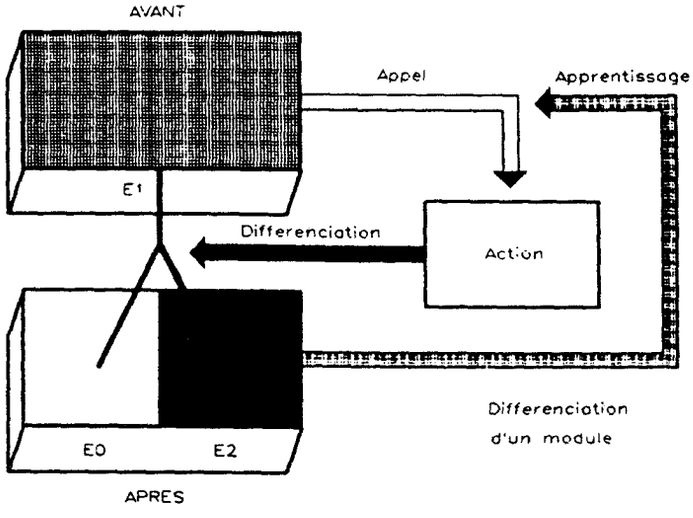


Figure 8. — Le mécanisme de division

5.3.3. Implantation du modèle

Au début de l'apprentissage, l'aire associative n'est pas différenciée. Lorsque le réseau ne reconnaît pas un phonème, on cherche tous les modules de l'aire associative qui ont développé une relation de déclenchement conditionnel [56] avec ce phonème. Tous ces modules sont alors divisés et l'apprentissage continue sur l'aire associative qui a toujours la même taille (un module se divise en deux modules de taille moitié) et qui possède des modules supplémentaires (un module divisé continue toujours à exister et peut, par exemple, déclencher inconditionnellement une réponse pour un autre phonème).

5.4. APPLICATION AU DÉCODAGE PHONÉTIQUE

5.4.1. Introduction

L'objectif que nous nous sommes fixé pour cette application du modèle de la colonne corticale concerne, d'une part, la reconnaissance de voyelles (aɛioyəø), d'autre part, la reconnaissance de fricatives (fvʃzs3̃). Nous ne pensons pas qu'un unique réseau puisse facilement traiter simultanément toutes les classes de phonèmes, les indices mis en cause pour la reconnaissance pouvant largement varier d'une classe à une autre (formants pour les voyelles, burst pour les occlusives, etc.). Notre premier travail consiste donc à élaborer un codage initial de l'information susceptible de mettre en valeur des indices spécifiques pour chaque classe de phonèmes.

5.4.2. Le codage initial

La théorie développée ici fournit une unité logique temporelle capable de distinguer des différences d'activité sur une carte bidimensionnelle d'entrée. Il va de soi que pour que l'automate et sa logique de fonctionnement puissent classifier des signaux, ceux-ci doivent être topologiquement séparables. C'est pourquoi il faut apporter un soin tout particulier dans le choix des indices représentés sur la carte d'entrée.

La représentation bidimensionnelle temps/fréquence de la parole engendre certains problèmes : débits variables, synchronisations de début, etc. qui sont autant d'obstacles à une invariance sur l'axe temporel indispensable à une bonne séparation des signaux. Le modèle de la colonne corticale étant intrinséquement parallèle, le temps n'est pas représenté sur la couche d'entrée, mais c'est plutôt par une succession temporelle d'entrées à laquelle correspond une succession de sorties, que le réseau gère le temps. Nous nous sommes tournés vers la neurobiologie pour spécifier un codage initial. Plusieurs remarques peuvent être faites :

— le prétraitement neuronal du signal de parole consiste, entre autres, en une décomposition fréquentielle,

— le système nerveux humain utilise souvent des différences locales d'activités [57],

— la dimension fréquentielle est représentée dans le cortex auditif sensoriel [59].

Nous avons par suite choisi d'effectuer une analyse cepstrale pour le prétraitement, à partir de laquelle nous devons extraire des indices primaires (représentés sur les axes de la carte) et des indices secondaires (variations locales pour un indice primaire donné).

5.4.2.1. *Codage des voyelles*

La reconnaissance des voyelles s'effectue typiquement sur des indices formantiques, dont la stabilité permet une bonne séparation ; le calcul des pics d'énergie nous fournit donc des possibilités intéressantes et justifiables biologiquement. Pour des distinctions plus fines (indices secondaires) nous utilisons les rapports d'énergies entre pics.

Tous ces indices sont recombinés sur la carte d'entrée en prenant comme première dimension les fréquences des pics inférieures à 1 300 Hz, comme deuxième dimension les fréquences des autres pics et localement le rapport d'énergie.

L'utilisation de $N - 2$ indices secondaires permet de représenter un espace de dimension N sur une carte bidimensionnelle [60].

5.4.2.2. *Codage des fricatives*

Les fricatives sont caractérisées par du bruit, en particulier en haute fréquence ; afin de simuler l'échelle logarithmique de l'oreille (Bark), nous effectuons un lissage fréquentiel très important à partir de 2 500 Hz, pour ensuite extraire des pics d'énergie et des pentes spectrales locales (justifiables biologiquement [61]).

Ces deux indices sont recombinés sur la carte d'entrée et ajustés localement par l'énergie des pics en indice secondaire (cf. voyelles).

5.4.3. Résultats expérimentaux

5.4.3.1. *Divisions*

Les figures 9 et 10, représentant les liens entre la sortie du système de décodage et l'aire associative, illustrent le phénomène de division.

Les trois premiers schémas de la figure 9 ont été obtenus par un apprentissage minimal sur un seul exemple de chaque fricative ; on constate bien sûr que la carte n'est pas très divisée : cinq niveaux de division ont suffi pour que le réseau sépare les trois différentes entrées.

Les trois derniers schémas de la figure 9 ont été réalisés sur un corpus d'apprentissage plus important : sept exemples pour chaque fricative. En de nombreux endroits, la division est maximale, c'est-à-dire égale à la résolution de la carte d'entrée.

Outre les niveaux de division successifs, ces six schémas font état des valeurs des coefficients probabilistes qui mémorisent l'apprentissage ; les relations de déclenchement inconditionnel (P2 proche de 1) sont représentées en noir, les relations d'inhibition (P2 proche de 0) en blanc et le déclenchement conditionnel est tramé. On peut ainsi suivre l'évolution des coefficients au cours des divisions successives des zones modulaires en sous-zones de rang inférieur.

Nous avons adopté la même représentation dans la figure 10 pour les voyelles.

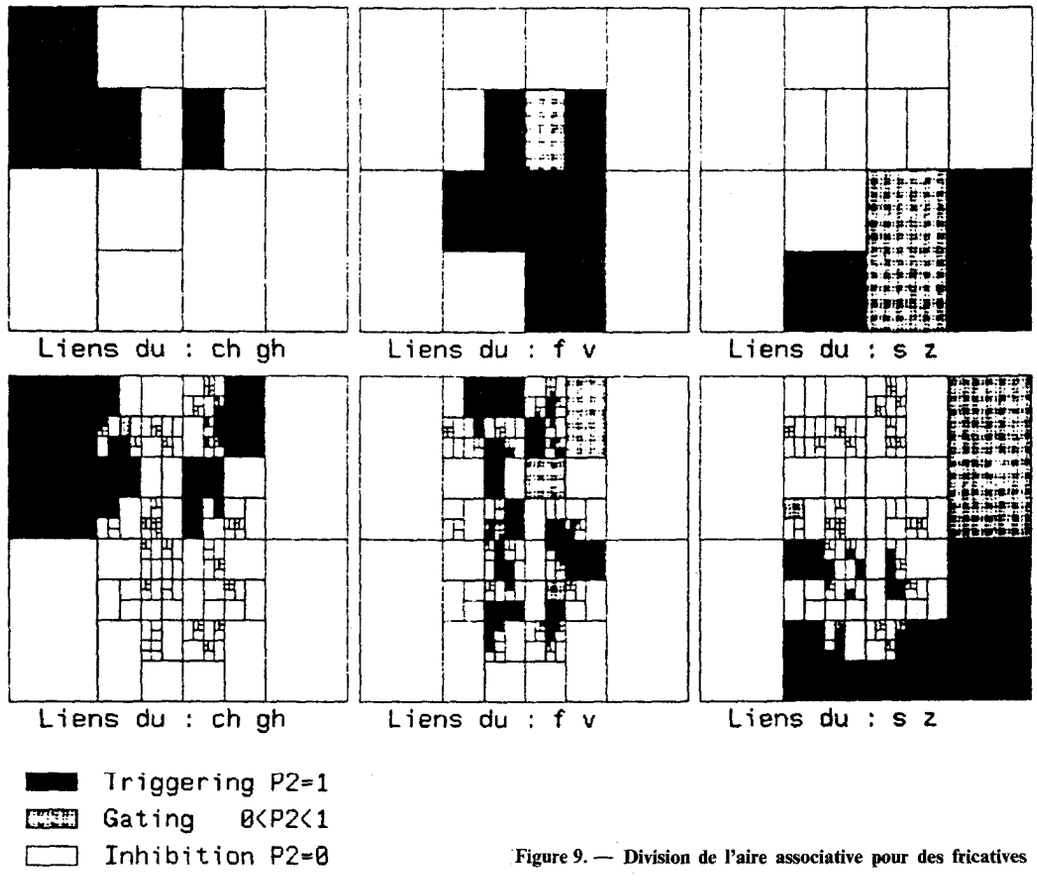


Figure 9. — Division de l'aire associative pour des fricatives

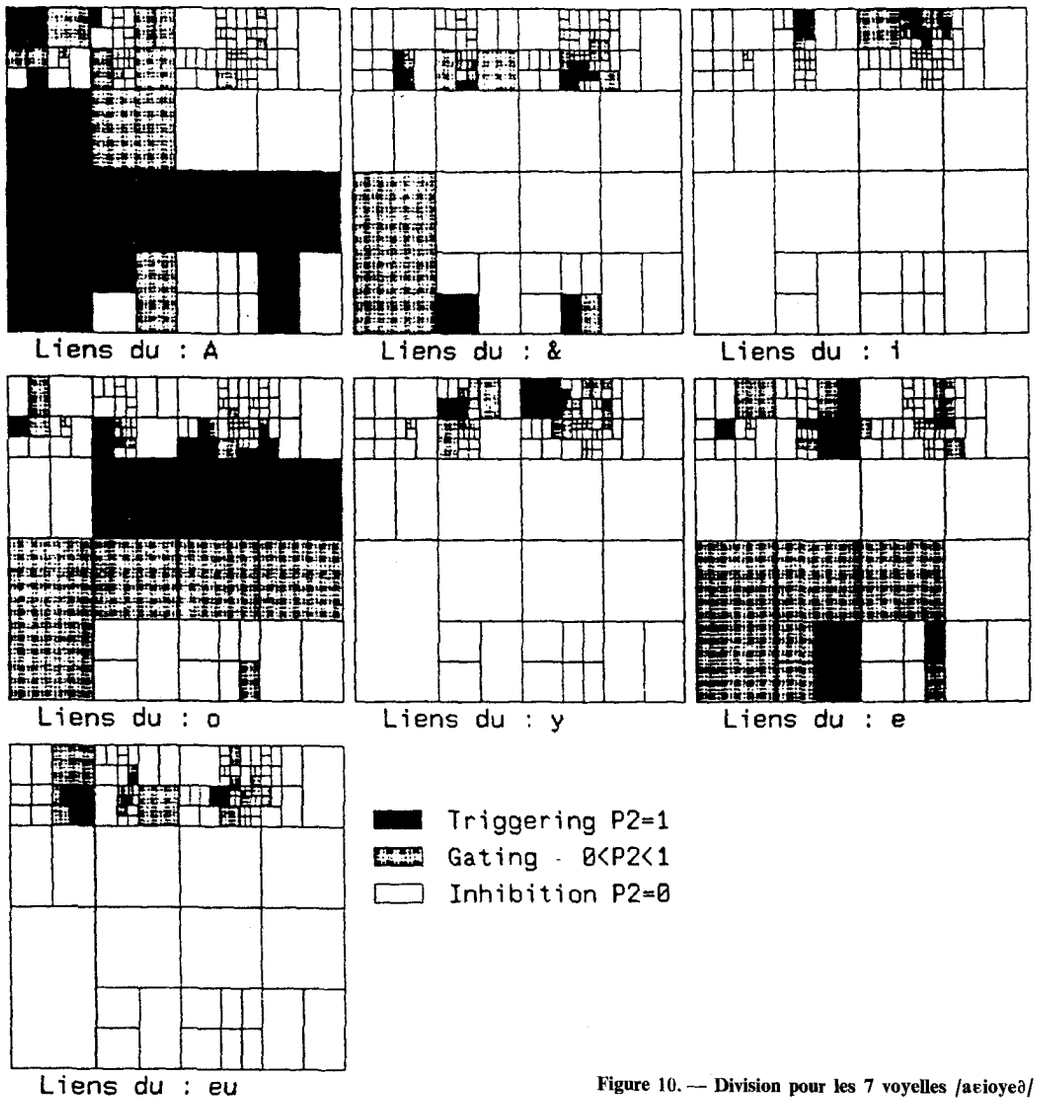


Figure 10. — Division pour les 7 voyelles /aɛioyeə/

5.4.3.2. *Performances*

Pour les fricatives, les corpus d'apprentissage sont constitués de phonèmes prononcés hors contexte par un seul locuteur. Les tests sont, en revanche, effectués en parole continue (25 mots contenant 50 fricatives) enregistrés sans précaution particulière. Le taux de reconnaissance pour un corpus d'apprentissage de 7 exemplaires par fricative atteint 96 %. Le réseau a exactement identifié 48 fricatives sur les 50 proposées en contexte et, de plus, n'a jamais proposé de solution sur les phonèmes des autres classes phonétiques.

Mais le plus intéressant réside dans le taux de reconnaissance atteint après apprentissage sur un seul exemple par fricative (pas de généralisation sur l'exemple possible). Bien qu'étant nettement moins divisé, le réseau reconnaît 47 fricatives sur le même corpus (94 %).

L'interprétation des taux de reconnaissance suggère quelques réflexions :

— le codage initial est satisfaisant tant par sa stabilité temporelle que par sa reproductibilité,

— trop de divisions successives conduisant à un grand nombre de zones peut nuire à la généralisation topologique. L'implantation d'un processus de fusion de zones serait souhaitable. Nous proposons, en accord avec la théorie biologique [57], un mécanisme possible de fusion locale de zones, fondé sur les relations de voisinage topologique. La reconnaissance s'effectuant essentiellement sur des signaux temporellement stables, les relations de déclenchement inconditionnel étant les plus significatives, ce processus de fusion, en stabilisant temporellement le module et en faisant croître les coefficients P2 de l'apprentissage, est tout-à-fait souhaitable pour des applications moins discriminantes.

En ce qui concerne les voyelles, l'apprentissage et les tests ont été effectués hors contexte sur 20 exemples par phonème pour l'apprentissage et 10 par phonème pour les

tests. On atteint un taux de reconnaissance global de 99 % pour 4 voyelles apprises et de 87 % pour 7 voyelles apprises. Il est important de noter que ces taux ne sont pas très significatifs sous cette forme et dépendent énormément des voyelles impliquées dans l'apprentissage. La proximité relative des différentes voyelles a bien été étudiée et le choix des voyelles à reconnaître influe sur les taux de reconnaissance de chaque voyelle. Ainsi, pour les 4 voyelles, nous avons choisi /aeio/. Leur distance relative explique le bon score global obtenu.

5.5. CONCLUSION

L'idée principale que nous avons exposée est sans aucun doute le principe de division. Cette nouvelle méthode d'apprentissage connexionniste fait converger le réseau vers une généralisation optimale tout en minimisant le nombre d'unités fonctionnelles.

Soulignons également la façon dont le réseau maîtrise la dimension temporelle d'un problème :

— entrée dynamique des données dans le réseau,
— sortie dynamique des réponses,
— rétro-action.

Notre approche fait évoluer le réseau dans le temps et non pas le temps dans le réseau.

Notre objectif, la reconnaissance de la parole continue, nous amènera à implanter le processus de fusion et le séquencement d'événements par arbres d'appel. Parallèlement, nous menons des travaux en reconnaissance et interprétation d'images utilisant la même modélisation.

6. Reconnaissance fondée sur la notion de centre des images acoustiques de symboles

6.1. PRINCIPE

Toutes les méthodes de segmentation de parole sont fondées sur l'hypothèse implicite que, représenté dans un certain espace paramétrique, le signal montre une transition entre deux unités phonétiques consécutives [62], [63]. Or, cette hypothèse n'est pas toujours vérifiée dans la parole continue ; lorsque la transition n'est pas nette, le positionnement précis de marques devient ambigu voire impossible. Cette situation apparaît fréquemment dans les phrases où la réalisation acoustique ne contient ni fricatives ni plosives. La segmentation reste donc difficile et pose souvent des problèmes de substitution, insertion et omission de symboles phonémiques dont la modélisation n'est pas encore satisfaisante. Pratiquement, des expériences ont montré que les résultats obtenus par des systèmes sans segmentation sont actuellement meilleurs que ceux avec segmentation [64], [65], [48]. Nous présentons ici une nouvelle formulation du problème de la reconnaissance de la parole continue sans segmentation, fondée sur la notion de centre d'images acoustiques des symboles phonétiques.

Notre méthode part du même principe que les méthodes de décomposition temporelle mais s'en distingue ensuite comme nous le montrons ci-dessous. La décomposition temporelle du signal donne une description de la parole en termes d'événements acoustiques que se recouvrent en partie. Chaque événement est caractérisé par une fonction d'interpolation à durée limitée et un coefficient spectral. Cette idée a été proposée initialement par Atal [78]. Elle a ensuite été reprise et améliorée, notamment pour la reconnaissance phonétique de la parole [79], [80]. Dans cette méthode, il est nécessaire d'effectuer une identification ultérieure des événements acoustiques pour les relier aux symboles phonétiques.

En revanche, la notion de centre d'images acoustiques des symboles que nous proposons permet de reconstruire une séquence de symboles phonétiques à partir d'une estimation de fonctions de plausibilité. Ces fonctions fournissent une mesure temporelle de la similarité entre le signal de parole et les images acoustiques des symboles phonétiques préalablement acquises. La reconnaissance est fondée sur un critère d'optimisation de la plausibilité globale.

En bref, les premiers travaux cités sont des *méthodes d'analyse* de la parole tandis que nous proposons une *méthode de reconnaissance.*

On considère que le signal de parole est une image acoustique de la séquence de symboles élémentaires composant la phrase. Par symbole élémentaire, nous entendons une unité élémentaire de décodage phonétique. L'hypothèse de base de cette approche est que l'image acoustique d'un symbole élémentaire est étirée selon l'axe du temps, son centre étant le plus influencé par le symbole. A chaque instant, le signal acoustique présente donc un certain degré de similarité avec les images de tous les symboles élémentaires possibles.

Soit :

A	: l'ensemble des symboles élémentaires,
$N = [0, T]$	: l'intervalle du signal vocal où T est le nombre d'échantillons du signal,
$M(a, n)$	: une fonction de $A \times N$ dans l'intervalle réel [0, 1] qui indique le degré de similarité du signal par rapport à l'image du symbole a à l'instant n,
s	: une phrase composée d'une suite finie d'éléments dans A : $s = a1, a2, ..., aL$ (s),
V	: un ensemble de M phrases différentes : $V = (s_1, s_2, ..., s_M)$.

Un signal de parole continue P*i* peut être vu comme une suite d'images acoustiques de symboles phonétiques *aj* constituant une phrase s_i dans V. La reconnaissance automatique de la parole continue fondée sur la notion de centre d'images acoustiques consiste à retrouver la suite des symboles produisant le signal à l'aide des deux étapes suivantes :

1) Le calcul d'une fonction de similarité $M(a, n)$ de $A \times N$ dans [0, 1], à partir de P*i*, qui mesure à l'instant n la vraisemblance du signal aux images d'un symbole a de A. Les maxima locaux de $M(a, n)$ en fonction de n sont appelés les centres d'images acoustiques du symbole a.

2) La recherche d'une suite de $L_{(s)}$ symboles a_k de telle sorte que l'image de ces symboles soit la plus cohérente possible avec le signal acoustique à reconnaître.

Nous détaillons ces étapes dans les deux paragraphes suivants.

6.2. FONCTION DE SIMILARITÉ

La fonction de similarité $M(a, n)$ indique le degré de similarité du signal avec les images acoustiques du symbole a à l'instant n.

Nous avons développé une méthode à base de comparaison de références acoustiques [66]. L'image du symbole a est modélisée par 2 à 5 profils échantillonnés uniformément à partir de la séquence de vecteurs du signal dans un certain espace paramétrique (LPCC, cepstre, Fourier...) :

$$I_a(m), \qquad 0 < m < M_a$$

où M_a est le nombre de profils dans la référence du symbole a. Ce modèle inclut des informations sur la durée moyenne, la variation maximale de durée et la position relative des événements du signal de l'image des symboles. Le calcul de $M(a, n)$ se fait en deux étapes :

1) *Calcul de la similarité instantanée* $R'_{a,n}$: A chaque instant, nous cherchons d'abord la meilleure similarité du signal P_n par rapport à tous les symboles, en faisant varier linéairement la durée de la référence :

$$R'_{a,n} = 1/M_a \max_{da1 < d < da2} L(P_{n-d/2}, I_a, M_a, d)$$

où

$$L(P_k, I_a, M_a, d) = \sum_{0 < j < Ma} 1(P_{k+jd/Ma}, I_a(j))$$

$da1$ et $da2$ sont respectivement la durée minimale et maximale de l'image du symbole a et $1(v_1, v_2)$ une mesure de vraisemblance locale entre les vecteurs v_1 et v_2. Cette mesure peut être le rapport de vraisemblance dérivé du LPCC, une transformation de la distance euclidienne, ou autre [67].

2) *Calcul de la similarité étendue* $M(a, n)$: Dans une deuxième étape, nous effectuons une extension de similarité sur $R'_{a,n}$, en utilisant l'information sur la durée des images des symboles. La durée de l'extension $D(a, n)$ est la durée optimale obtenue pendant l'étape précédente :

$$M(a, n) = \max_{-D(a,n)/2 < k < D(a,n)/2} R'_{a,n} \times W(k, D(a, n))$$

avec la fonction de pondération :

$$W(k, d) = 1 - |k| \times 2(1 - MINV)/d, \quad |k| < d/2$$

MINV est une constante entre 0 et 1 ; une valeur de 0,7 a donné des résultats satisfaisants.

Afin de rendre compte des déformations contextuelles d'articulation, nous utilisons des références multiples. Les références sont codées par quantification vectorielle pour réduire les calculs. Les aspects techniques de l'algorithme sont détaillés dans [68].

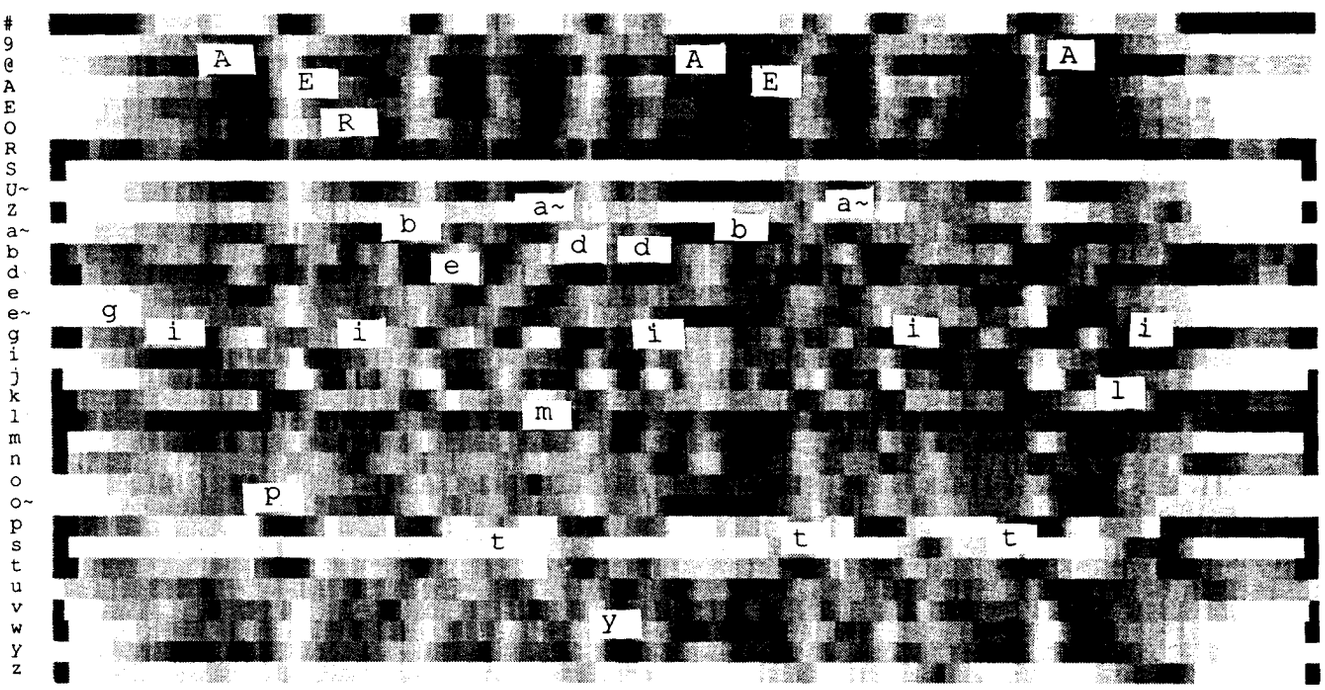


Figure 11. — Fonction de similarité $M(a, n)$ obtenue pour la phrase : « Guy a péri bêtement du diabète en Italie » (locuteur masculin). L'axe horizontal correspond au temps. L'axe vertical correspond aux symboles phonétiques, dont le nombre est de 32, triés alphabétiquement. Le niveau de gris correspond à la valeur de $M(a, n)$, noir représentant 1 et blanc 0. L'étiquetage des symboles phonétiques est fait manuellement pour illustrer la correspondance entre les symboles et leur image.

Sur la figure 11 nous donnons un exemple de représentation de la fonction de vraisemblance $M(a, n)$ obtenue pour une phrase.

Toutes les méthodes capables de fournir la fonction de similarité peuvent être employées. Celles à base de modèles neuroniques [69], [70], [71] sont particulièrement adaptées. Les résultats des autres méthodes fondées sur des systèmes à bases de connaissances, cf. par exemple APHODEX [72] ou [73], ou sur des modèles markoviens cachés [65], [74] sont aussi utilisables moyennant quelques modifications.

6.3. RECHERCHE DE LA MEILLEURE SÉQUENCE DE SYMBOLES

La recherche de la meilleure séquence de symboles consiste à trouver une suite de transitions temporelles $t(ak)$ en fonction de $M(a, n)$ qui, sous contraintes syntaxiques, lexicales, phonologiques et phonétiques, maximisent la quantité :

$$Q(s) = \sum_{k=0}^{L(s)-1} \sum_{n=t(a_k)}^{t(a_{k+1})-1} M(a_k, n) .$$

Cette recherche est guidée par les maxima en fonction de n dans $M(a, n)$. La solution de ce problème utilise deux formalismes de programmation dynamique [75] pour fournir le mot S_r de meilleur score parmi l'ensemble des phrases :

$$S_r = \text{argmax}_s\, Q(s) .$$

Le premier calcule les meilleurs chemins liés aux centres d'images des symboles consécutifs et le deuxième calcule les meilleures transitions temporelles.

6.4. CONCLUSION

Cette approche du décodage phonétique permet une meilleure utilisation des centres d'images acoustiques des symboles phonétiques peu influencés par les effets de contexte, par rapport aux frontières de ces images. Grâce à cette propriété, la taille de corpus d'apprentissage pour un locuteur peut être réduite à environ 20 phrases, ce qui est peu par rapport aux autres méthodes de reconnaissance.

Les problèmes introduits par la segmentation, tels que l'insertion, l'omission et la substitution des unités élémentaires de reconnaissance, sont ainsi évités. Les taux de reconnaissance que nous avons obtenus sont comparables aux meilleurs résultats obtenus par les autres systèmes [76], [77].

7. Bilan et conclusions

Malgré les grands progrès effectués dans la reconnaissance de la parole par méthodes globales ou semi-globales (notamment à l'aide de modèles stochastiques markoviens), le décodage acoustico-phonétique de la parole continue (DAP) demeure une étape obligée dans la conception de systèmes avancés d'interaction orale homme-machine.

Nous avons présenté dans cet article la problématique du DAP et ses difficultés en illustrant notre propos par la description des diverses approches complémentaires adoptées par notre équipe pour progresser vers la solution de ce problème.

L'utilisation explicite de connaissances acoustiques et phonétiques, fondée sur des techniques de systèmes à bases de connaissances, permet d'atteindre des performances notables, en particulier en reconnaissance indépendante du locuteur. Ceci est attesté par les résultats obtenus grâce au système expert APHODEX que nous avons développé en nous fondant au départ sur les connaissances d'un expert phonéticien en lecture de spectrogrammes vocaux. Ce système utilise le phonème comme unité de segmentation et de reconnaissance. Cela introduit des problèmes liés à la grande variabilité contextuelle de cette unité. Pour pallier ces difficultés, nous avons utilisé le triplet phonétique comme unité de base. Cette unité présente l'avantage d'inclure les variations phonologiques aux frontières. Elle soulève en revanche d'autres questions liées, d'une part, au grand nombre de triplets nécessaires pour représenter correctement une langue telle que le français et à l'acquisition de la base de connaissances correspondante et, d'autre part, à la prise en compte des contraintes linguistiques que doit respecter une séquence de triplets et à la propagation de ces contraintes le long d'un treillis phonétique.

La conjonction des techniques de l'intelligence artificielle et d'une unité de reconnaissance contextuelle telle que le triplet phonétique semble être ainsi très prometteuse. Néanmoins, des problèmes importants demeurent, en particulier pour la représentation et l'identification des unités individuelles. Les modèles connexionnistes, fondés sur une approximation plus ou moins fidèle de la réalité neuro-biologique du cerveau humain constituent une voie intéressante pour améliorer cet aspect. Nous avons également présenté dans cet article un modèle connexionniste original fondé sur la notion de colonne corticale et nous avons montré sur deux exemples (la reconnaissance de voyelles et des consonnes fricatives) que ce modèle possédait de bonnes propriétés d'apprentissage et de différenciation de formes acoustiques. On peut penser que de tels modèles constiueront un ingrédient parmi d'autres des systèmes de DAP à venir.

Enfin, un problème constant en DAP, source de multiples erreurs, est celui de la segmentation du continuum acoustique en unités élémentaires. Une solution à ce problème consiste à différer le plus longtemps possible la segmentation. Nous présentons ainsi une méthode de reconnaissance sans segmentation fondée sur l'étude des variations des « images acoustiques » des unités phonémiques utilisées. Cette méthode met en œuvre des techniques de recherche de chemin optimal par programmation dynamique grâce aux variations d'une fonction de similarité. Elle fournit de bons résultats et présente l'avantage de pouvoir être couplée à d'autres méthodes de calcul de similarité

entre formes (modèles markoviens, neuronaux ou à bases de connaissances explicites).

En conclusion, le DAP est un problème difficile et multiforme qui nécessite une approche multiple fondée sur des méthodes complémentaires. Les diverses méthodes développées par notre équipe et présentées dans cet article font à notre avis partie des méthodes prometteuses pour l'avenir.

REMERCIEMENTS

Cet article présente les travaux menés au sein de l'équipe « Reconnaissance des Formes et Intelligence Artificielle » du CRIN/INRIA-Nancy, dans le domaine du décodage acoustico-phonétique. Outre les auteurs, les personnes dont les noms suivent ont contribué à ces travaux et sont chaleureusement remerciées : Frédéric Alexandre, Noëlle Carbonell, Catherine Dingeon, Mahieddine Djoudi, Dominique François, Frédéric Guyot, Marie-Christine Haton, François Lonchamp, Odile Mella.

Nous remercions également les lecteurs de cet article pour leurs remarques constructives.

Manuscrit reçu le 21 novembre 1989.

BIBLIOGRAPHIE

[1] W. A. Lea, *Trends in Speech Recognition,* Prentice-Hall, 1980.

[2] J. P. Haton, *Intelligence artificielle en compréhension automatique de la parole : état des recherches et comparaison avec la vision par ordinateur,* TSI, vol. 4, n° 3, 1985, pp. 265-287.

[3] D. Klatt, *Review of the ARPA Speech Understanding Project,* JASA, 62, 1977, pp. 1345-1366.

[4] F. Lonchamp, *Reading Spectrograms : the View of the Expert,* in « Fundamentals in Computer Understanding : Speech and Vision », J. P. Haton, editor, Cambridge University Press, 1987.

[5] J. Caelen, N. Vigouroux, G. Pérennou, *Structuration des informations acoustiques dans le projet ARIAL,* Speech Com., vol. 2, 2-3, 1983, pp. 219-222.

[6] J. P. Haton, *Contribution à l'analyse, la paramétrisation et la reconnaissance automatique de la parole,* Thèse de Doctorat d'État, Université de Nancy I, 1974.

[7] H. Meloni, *Étude et réalisation d'un système de reconnaissance automatique de la parole,* Thèse de Doctorat d'État, Université de Marseille, 1982.

[8] J. S. Liénard, *Analyse, synthèse et reconnaissance automatique de la parole,* Thèse de Doctorat d'État, Université de Paris 6, 1972.

[9] C. Scagliola, *Continuous Speech Recognition without Segmentation : Two Ways using Diphones as basic Speech Units,* Speech Com., vol. 2, 2-3, 1983, pp. 199-201.

[10] G. Ruske, *Automatic Recognition of Syllabic Speech Segments Using Spectral and Temporal Features,* Proc. ICASSP-82, Paris, 1982.

[11] O. Fujimura, *Syllabes as Concatened Demi-Syllabes and Affixes,* 9th Meeting ASA, 1976.

[12] G. Pérennou, *The ARIAL II Speech Recognition System,* in « Automatic Speech Analysis and Recognition », J. P. Haton, editor, D. Reidel Publishing Company, 1982.

[13] R. de Mori, *Extraction of Acoustic Cues Using a Grammar of Frames,* Speech Com., vol. 2, 2-3, 1983, pp. 223-225.

[14] G. Mercier, *The KEAL Speech Understanding System,* in « Spoken Language Generation and Understanding », J. C. Simon, editor, D. Reidel, 1983.

[15] D. Fohr, J. P. Haton, F. Lonchamp, L. Sauter, *Méthodes de segmentation syllabique en reconnaissance de la parole,* 14èmes JEP, GALF, Paris, 1985.

[16] R. M. Schwartz *et al., Improved Hidden Markov Modeling of Phonemes for Continuous Speech Recognition,* Proc. Int. Conf. ASSP, 1984.

[17] J. M. Baker, *Stochastic Modelling for Automatic Speech Understanding,* in « Speech Recognition », R. Reddy, editor, New York, Academic Press, 1975.

[18] H. G. Goldberg, *Segmentation and Labelling of Speech : A Comparative Performance Evaluation,* Ph. D. Thesis, Carnegie-Mellon University, 1975.

[19] T. G. von Keller, *An On-line Recognition System for Spoken Digits,* JASA, 49, 1971, pp. 1288-1296.

[20] M. Baudry, *Étude du signal vocal dans sa représentation amplitude-temps. Algorithme de segmentation et de reconnaissance de parole,* Thèse de Doctorat d'État, Université de Paris 6, 1978.

[21] R. de Mori, G. Giordano, *A Parser for Segmenting Continuous Speech into Pseudo-Syllabic Nuclei,* Proc. ICASSP-80, 1980.

[22] R. Mizoguchi, O. Kakusko, *Continuous Speech Recognition Based on Knowledge Engineering Techniques,* Proc. ICASSP-84, 1984.

[23] S. Roucos *et al., Vector Quantization for Very-Low-Rate,* Proc. Global Telecom. Conf., Miami, 1982.

[24] F. Jelinek, *Continuous Speech Recognition by Statistical Methods,* Proc. IEEE, vol. 64, 4, 1976.

[25] M. Cravero, R. Pieraccini, F. Raineri, *Definition and Evaluation of Phonetic Units for Speech Recognition by Hidden Markov Models,* Proc. ICASSP-86, Tokyo, 1985.

[26] S. R. Levinson, L. R. Rabiner, M. M. Sondhi, *An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition,* The Bell System Technical Journal, vol. 62, 4, 1983, pp. 1035-1074.

[27] P. D. Green *et al., A Speech Recognition Strategy Based on Making Acoustic Evidence and Phonetic Knowledge Explicit,* Proc. European Conf. Speech Technology, Edinburg, 1987.

[28] W. Lea, *Trends in Speech Recognition,* Prentice-Hall, 1983.

[29] R. Mizoguchi, O. Kakusho, *Continuous Speech Recognition Based on Knowledge Engineering Techniques,* Proc. ICASSP-84, 1984, pp. 638-640.

[30] D. Memmi, M. Eskénazi, J. Mariani, A. Nguyen-Xuan, *Un système expert pour la lecture de sonagrammes,* Speech Com., vol. 2, n° 2-3, 1983, pp. 234-236.

[31] N. Carbonell, M. O. Cordier, D. Fohr, J. P. Haton, F. Lonchamp, J. M. Pierrel, *Acquisition et formalisation du raisonnement dans un système expert de lecture de spectrogrammes vocaux,* Actes du colloque ARC (Association de la Recherche Cognitive), « Les modes de raisonnement », Paris, avril 1984.

[32] J. P. Haton, N. Carbonell, D. Fohr, J. F. Mari, A. Kriouile, *Interaction between Stochastic Modeling and Knowledge-Based Techniques in Acoustic-Phonetic Decoding of Speech,* Proc. ICASSP-87, Dallas, April 1987.

[33] N. Carbonell, D. Fohr, J. P. Haton, *APHODEX, An Acoustic-Phonetic Decoding Expert System,* International Journal of Pattern Recognition and Artificial Intelligence, vol. 1, n° 2, 1987, pp. 207-222.

[34] G. Fant, *A Note on Vocal Tract Size Factors and Non-Uniform F-Pattern scalings,* Speech Sounds and Features, The MIT Press, 1973, pp. 84-93.

[35] H. Traünmuller, *Articulatory and Perceptual Factors Controlling the Age and Sex-Conditioned Variability in Formant Frequencies of Vowels,* Speech Com., vol. 3-1, 1984, pp. 49-62.

[36] J. P. Tubach, L. J. Boë, *Un corpus de transcriptions phonétiques : constitution et exploitation statistique,* Rapport ENST-85D001, avril 1985.

[37] D. Fohr et Y. Laprie, *SNORRI : An Interactive System for Speech Analysis,* Proc. of The European Conference on Speech Technology, Paris, septembre 1989.

[38] O. Mella, M. C. Haton, *Méthodologie d'étude de la pertinence de paramètres phonétiques et acoustiques pour la reconnaissance du locuteur,* Actes du Séminaire sur la variabilité du locuteur, Luminy, juin 1989.

[39] R. S. Michalski, R. E. Stepp, *Learning from Observation : Conceptual Clustering,* chapitre 11, « Machine Learning : An Artificial Intelligence Approach », R. S. Michalski, J. G. Carbonnell and T. M. Mitchell Ed., Springer Verlag, 1984.

[40] A. Bonneau, D. Fohr, *Normalisation du locuteur par modélisation implicite de la longueur du conduit vocal,* Actes du Séminaire sur la variabilité du locuteur, Luminy, juin 1989.

[41] O. Engstrand, *Articulatory Correlates of Stress and Speaking Rate in Swedish VCV Utterances,* J. Acoust. Soc. Am., vol. 85 (5), 1988, pp. 1863-1875.

[42] M. Minsky, *A Framework for Representing Knowledge,* in « The Psychology of Computer Vision », P. H. Winston, editor, Mc Graw-Hill Book Company, 1975.

[43] C. Granger, *Reconnaissance d'objets par mise en correspondance en vision par ordinateur,* Thèse de Doctorat, Nice, 1985.

[44] J. P. Damestoy, *Réalisation d'un système à base de prototypes pour le contrôle du décodage acoustico-phonétique de la parole,* Thèse de l'Université de Nancy 1, 1986.

[45] P. D. Green, M. P. Cooke, H. H. Lafferty, A. J. H. Simons, *A Speech Recognition Strategy Based on Making Acoustic Evidence Knowledge Explicit* in « Recent Advances in Speech Understanding and Dialog Systems », H. Niemann, M. Lang and G. Sagerer Ed., NATO ASI Series, Springer-Verlag, 1988.

[46] D. Waltz, *Understanding Line Drawings of Scenes with Shadows,* in « The Psychology of Computer Vision », P. H. Winston, editor, Mc Graw-Hill, New York, 1975.

[47] R. Mohr, G. Masini, *Good Old Discrete Relaxation,* Proc. ECAI, Munich, 1988, pp. 651-656.

[48] K. F. Lee, Hsiao-Wuen Hon, Mei-Yyuh Hwang, S. Mahajan, R. Reddy, *The SPHINX Speech Recognition System,* Proc. ICASSP-89, Glasgow, Scotland, 1989.

[49] J. C. Platt, J. J. Hopfield, *Analog Decoding Using Neural Networks,* AIP Conf. Neural Networks for Computing, Snowbird, J. S. Denker, editor, 1987.

[50] I. S. Howard, M. A. Huckvale, *Speech Fundamental Period Estimation Using a Trainable Pattern Classifier,* FASE Speech 88, Edinburgh, 1988.

[51] F. Yang, L. Wu, J. P. Haton, *Utilisation d'un réseau de neurones pour la reconnaissance des mots isolés,* FASE Speech 88, Edinburgh, 1988.

[52] H. Bourlard, C. J. Wellekens, *Speech Dynamics and Recurrent Neural Networks,* Proc. ICASSP-89, Glasgow, Scotland, 1989.

[53] H. Sakoe, R. Isotani, K. Yoshida, K. Iso, T. Watanabe, *Speaker Independent Word Recognition Using Dynamic Programming Neural Networks,* Proc. ICASSP-89, Glasgow, Scotland, 1989.

[54] A. Waibel, H. Sawai, K. Shikano, *Consonant Recognition by Modular Construction of Large Phonemic Time Delay Neural Networks,* Proc. ICASSP-89, Glasgow, Scotland, 1989.

[55] F. Guyot, F. Alexandre, J. P. Haton, *Toward a Continuous Model of the Cortical Column : Application to Speech Recognition,* Proc. ICASSP-89, Glasgow, Scotland, 1989.

[56] F. Alexandre, Y. Burnod, F. Guyot, J. P. Haton, *La colonne corticale : nouvelle unité de base pour des réseaux multicouches,* Proc. Neuro-Nîmes' 89, Nîmes, 1989.

[57] Y. Burnod, *An Adaptative Neural Network : the Cerebral Cortex,* Masson, Paris, 1988.

[58] F. Guyot, F. Alexandre, J. P. Haton, Y. Burnod, *The Cortical Column, a New Processing Unit for Cortex-Like Networks,* Proc. CEC Workshop « From the Pixels to the Features », Elsevier, 1989.

[59] E. I. Knudsen, S. du Lac, S. D. Esterly, *Computational Maps in the Brain,* Neurosciences, 10, 1987, pp. 41-65.

[60] D. H. Ballard, *Cortical Connections and Parallel Processing : Structure and Function,* The Behavioral and Brain Sciences 9, 1986, pp. 67-120.

[61] B. Delgutte, *Codage de la parole dans le nerf auditif,* Thèse de l'Université de Paris 6, 1984.

[62] R. Zelinski, *A Segmentation Algorithm for Connected Word Recognition Based on Estimation Principles,* IEEE Trans. Acoust., Speech and Signal Processing, ASSP-31, 1983, pp. 818-827.

[63] P. Mermelstein, *Automatic Segmentation of Speech into Syllabic Units,* J. Acoust. Soc. Am., 58, 1975, pp. 880-883.

[64] C. Myers, L. R. Rabiner, *A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition,* IEEE Trans. Acoust., Speech and Signal Processing, ASSP-29 (2), 1981, p. 284.

[65] L. R. Bahl, F. Jelinek, R. L. Mercer, *A Maximum Likelihood Approach to Continuous Speech Recognition,* IEEE Trans. PAMI, PAMI-5 (2), 1983, pp. 179-190.

[66] Y. Gong, J. P. Haton, *Phoneme Based Continuous Speech Recognition without Pre-segmentation,* Proc. of European Conference on Speech Technology, Edinburgh, September 1987, pp. 121-124.

[67] N. Nocerino, F. K. Soong, L. R. Rabiner, D. H. Klatt, *Comparative Study of Several Distorsion Measures for Speech Recognition,* Proc. IEEE ICASSP-85, 1985, pp. 25-28.

[68] Y. Gong, *Contribution à l'interprétation automatique des signaux en présence d'incertitude,* Thèse de l'Université de Nancy I, mai 1988.

[69] A. J. Robinson, F. Fallside, *A Dynamic Connectionist Model for Phoneme Recognition,* in Euro'88, ENST, Paris, France, June 1988.

[70] M. A. Franzini, M. J. Witbrock, K. F. Lee, *A Connectionist Approach to Continuous Speech Recognition,* in Proc. ICASSP-89, Glasgow, Scotland, May 1989.

[71] H. Sawai, A. Waibel, M. Miyatake, K. Shikano, *Spotting Japanese CV-syllabes and Phonemes using Time-Delay Neural Networks,* Proc. ICASSP-89, Glasgow, Scotland, May 1989.

[72] N. Carbonell, J. P. Haton, D. Fohr, F. Lonchamp, J. M. Pierrel, *APHODEX, Design and Implementation of an Acoustic-phonetic Decoding Expert System,* Proc. ICASSP-86, Tokyo, Japan, 1986.

[73] Y. Gong, J. P. Haton, *A Knowledge Based System for Contextually Deformed Pattern Interpretation Applied to Chinese Tone Recognition,* Proc. of 2nd International Conference on Artificial Intelligence, IRIAM, Marseille, 1986, pp. 521-530.

[74] J. K. Baker, *Stochastic Modeling for Automatic Speech Understanding,* in « Speech Recognition », D. R. Reddy, editor, Academic Press, 1975.

[75] Y. Gong, J. P. Haton, *Signal-to-String Conversion Based on High Likelihood Regions Using Embedded Dynamic Programming,* IEEE Trans. PAMI, to appear.

[76] Y. Gong, F. Mouria, J. P. Haton, *Un système de reconnaissance de la parole continue sans segmentation,* Actes du 7ème Congrès AFCET « Reconnaissance des Formes et Intelligence Artificielle, Paris, novembre 1989.

[77] J. MARIANI, *Recent Advances in Speech Processing,* Proc. ICASSP-89, Glasgow, Scotland, May 1989.

[78] B. S. ATAL, *Efficient coding of LPC parameters by temporal decomposition,* in Proc. ICASSP-83, pp. 81-84.

[79] F. BIMBOT, G. CHOLLET, P. DELEGLISE and C. MONTACIÉ, *Temporal decomposition and acoustic-phonetic decoding of speech,* in Proc. ICASSP-88, pp. 445-448, New York, USA, 1988.

[80] G. BAILLY, P. F. MARTEAU, C. ABRY, *A new algorithm for temporal decomposition of speech. Application to a numerical model of coarticulation,* in Proc. ICASSP-89, pp. 508-511, Glasgow, Scotland, 1989.