Université de Nancy I

U.E.R. Sciences Mathématiques

Centre de Recherche en Informatique de Nancy

MIMULE: UN SYSTEME DE RECONNAISSANCE DE MOTS ISOLES MULTILOCUTEURS UTILISANT LES TECHNIQUES DE CLASSIFICATION

THESE

présentée et soutenue publiquement le 31 mars 1988

A L'UNIVERSITE DE NANCY I

pour l'obtention du titre de

DOCTEUR D'UNIVERSITE EN INFORMATIQUE

par

Pascal DIVOUX

Composition du Jury

Président:	Jean-Marie PIERREL
Rapporteurs:	Laurent MICLET
	Jean-Laurent MALLET
Examinateurs:	Marie-Christine HATON
	Jean-Paul HATON

Université de Nancy I U.E.R. Sciences Mathématiques

Centre de Recherche en Informatique de Nancy

MIMULE: UN SYSTEME DE RECONNAISSANCE DE MOTS ISOLES MULTILOCUTEURS UTILISANT LES TECHNIQUES DE CLASSIFICATION

THESE

présentée et soutenue publiquement le 31 mars 1988

A L'UNIVERSITE DE NANCY I

pour l'obtention du titre de

DOCTEUR D'UNIVERSITE EN INFORMATIQUE

par

Pascal DIVOUX

Composition du Jury

Président:	Jean-Marie PIERREL
Rapporteurs:	Laurent MICLET
	Jean-Laurent MALLET
Examinateurs:	Marie-Christine HATON
	Jean-Paul HATON

Je tiens à exprimer ma profonde gratitude à Jean-Paul Haton, qui a bien voulu m'accueillir dans son laboratoire et assurer la direction de cette thèse.

Que Jean-Marie Pierrel trouve ici l'expression de ma reconnaissance pour avoir bien voulu présider à ce jury.

Messieurs les professeurs Laurent Miclet et Jean-Laurent Mallet me font l'honneur de juger ce travail, qu'ils en soient vivement remerciés.

Je tiens à remercier également Marie-Christine Haton pour l'intéret amical qu'elle porte à mes travaux.

Que tous les membres de l'équipe RFIA, et particulièrement Anne et Yifan, soient remerciés pour leurs conseils amicaux et le temps qu'ils m'ont consacré.

Merci enfin à tous ceux qui, par leur aide matérielle, morale, financière ou logistique ont largement participé à la réalisation de ce travail.

PLAN

Partie A: La reconnaissance de la parole

Partie B : Le système MIMULE

PARTIE A

LA RECONNAISSANCE DE LA PAROLE

CHAPITRE 1

INTRODUCTION

La généralisation sans cesse croissante de l'outil informatique et sa mise à disposition du plus large public rendent de plus en plus pressantes la mise au point et l'amélioration de nouvelles techniques de communication entre l'homme et l'ordinateur. D'ores et déjà souris et icônes tendent à supplanter dans de nombreux cas les traditionnels écrans-claviers; cependant, l'homme communiquant avec la machine est toujours sollicité suivant les mêmes ''canaux'' : Les yeux (pour la perception) et les mains (pour l'action).

Il est incontestable qu'un poste de travail muni d'un périphérique de dialogue oral (synthèse vocale et reconnaissance de la parole) présenterait de nombreux avantages tant fonctionnels qu'ergonomiques:

- La parole est, de tous nos moyens de communication, celui qui nous est le plus naturel, le plus spontané et le plus adapté à l'interactivité.

- Sollicitant un ''canal'' différent (voix et ouie), la parole permet l'utilisation d'une machine dans des contextes où celà n'est pas possible habituellement: opérateur ayant les mains ou/et les yeux occupés, (microchirurgie, conduite de machine-outil, de véhicule); utilisation par des handicapés moteurs ou de la vue...

- De plus, les ondes sonores sont omnidirectionnelles et peu influencées par les obstacles, ce qui permet à l'opérateur de maintenir le contact lors d'un déplacement à faible échelle.

Ces propriétés intéressantes du dialogue oral homme-machine sont également responsables de leur principal inconvénient: Ce mode de communication est beaucoup moins fiable que les modes traditionnels car fugace, généralement plus bruité et très variable.

C'est cette variabilité (inter- et intralocuteur) du signal vocal qui fait la difficulté principale des systèmes de reconnaissance de la parole. Certaines applications peuvent se satisfaire d'une restriction à un seul

utilisateur afin de la minimiser: Utilisation d'objets personnels, identification du locuteur par empreintes vocales... Toutefois, de nombreuses applications doivent être à la portée d'un grand nombre de personnes (centre de renseignements, banque de données, manipulation d'outils mis à la disposition d'une collectivité, etc...); c'est là le but des recherches en reconnaissance de la parole multilocuteur.

C'est dans ce cadre que s'insère le travail que nous présentons ici: Un système de reconnaissance de la parole multilocuteur fonctionnant dans un contexte simple : La simulation de reconnaissance vocale de codes chiffrés (vocabulaire limité et prononciation en mots isolés).

La partie A présente les généralités sur la reconnaissance de la parole en se concentrant sur les applications multilocuteurs.

Le chapitre 2 décrit l'appareil vocal, les mécanismes de production des sons et leur influence sur les caractéristiques du signal de parole.

Le chapitre 3 présente les différentes applications de la reconnaissance de la parole ainsi que les types de stratégies utilisées.

Le chapitre 4 détaille les différentes méthodes actuelles de gestion de la variabilité dans les systèmes de reconnaissance multilocuteurs.

Le chapitre 5 décrit quelques méthodes de classification automatique citées dans les chapitres précédents.

La partie B est consacrée à la description du système MIMULE qui développe certaines idées décrites en partie A (Comparaison dynamique globale, mots isolés, contexte multilocuteur) autour d'une application de reconnaissance automatique de codes chiffrés.

Les 4 premiers chapitres présentent le contexte et la structure de base constante du projet; le chapitre 5 est consacré aux caractéristiques essentielles des locutions du corpus-test; le chapitre 6 détaille les différents choix de variantes à partir de la structure de base ainsi que leur influence sur les taux de reconnaissance et le chapitre 7 analyse et discute les résultats obtenus.

CHAPITRE 2

LA PAROLE

21. INTRODUCTION

Nous abordons dans ce chapitre les mécanismes de production de la parole, et leurs caractéristiques principales. Notre but n'est pas d'en faire une description détaillée, mais d'en donner un aperçu qui permette de comprendre mieux l'origine des difficultés spécifiques de la reconnaissance automatique multilocuteur de la parole. Pour une description plus complète et précise on pourra consulter par exemple [Flanagan 65], [Malmberg 71], [Carton 74].

Après une présentation schématique de l'appareil vocal (paragraphe 22), nous abordons les mécanismes mis en jeu par celui-ci lors de la production des sons : voyelles et consonnes (paragraphe 23). Les deux paragraphes suivants traitent de deux caractéristiques importantes du signal vocal dont il faut tenir compte dans tout système de reconnaissance de la parole : la variabilité (paragraphe 24) et la redondance (paragraphe 25). Nous terminons par un aperçu des différentes méthodes de codage du signal vocal (paragraphe 26).

Remarque: Les symboles utilisés dans ce chapitre pour désigner les phonèmes sont ceux de l'alphabet phonétique international (A.P.I.), la figure A1 les récapitule et en donne un exemple dans un mot français.

Phonème	Mot-clef	Classe	Phonème	Mot-clef	Classe
a	patte	Voyelles orales	p	passe	Plosives
ɑ	pâte		t	toux	
i	riz		k	cou	
y	rue		b	basse	
ɔ	port		d	doux	
o	pot		g	goût	
ə	le				
ɛ	raie		m	masse	Nasales
e	ré		n	nous	
ø	peu		ɲ	signe	
œ	peur				
u	roue		f	fer	Fricatives
			s	assis	
ɑ̃	blanc	Voyelles nasales	ʃ	chou	
ɔ̃	bon		v	verre	
ɛ̃	vin		z	Asie	
œ̃	un		ʒ	joue	
j	hier	Semi-consonnes	l	la	Liquides
ɥ	huit		R	rat	
w	oui				

Figure A1: Les phonèmes du français d'après [Carton 74]

22. DESCRIPTION DE L'APPAREIL VOCAL

L'appareil vocal humain comprend les poumons, le larynx et le conduit vocal (pharynx, bouche et nez) cf. Figure A2.

• Les poumons constituent la source principale d'énergie dans la production des sons, leur rôle est celui d'un soufflet qui créé un flux d'air plus ou moins puissant déterminant ainsi l'intensité de la voix (voix basse, forte, cris...).

• Le larynx est le lieu de production de l'onde glottique c'est-à-dire de la partie périodique des sons voisés (fondamental). Cette onde est produite par la vibration des cordes vocales (muscles élastiques du larynx) au passage du flux pulmonaire. La tension de celles-ci associée à la pression sous-glottique détermine la hauteur du son produit, c'est-à-dire sa fréquence. (voix grave ou aigue).

• Le conduit vocal (pharynx, bouche, cavité nasale) est la partie de l'appareil phonatoire qui imprime au son émis les caractéristiques spécifiques permettant de distinguer les divers phonèmes et ceci selon deux fonctions différentes :

- En tant que résonateur de l'onde glottique pour la production des voyelles.
- En tant que générateur de bruit pour la production des consonnes.

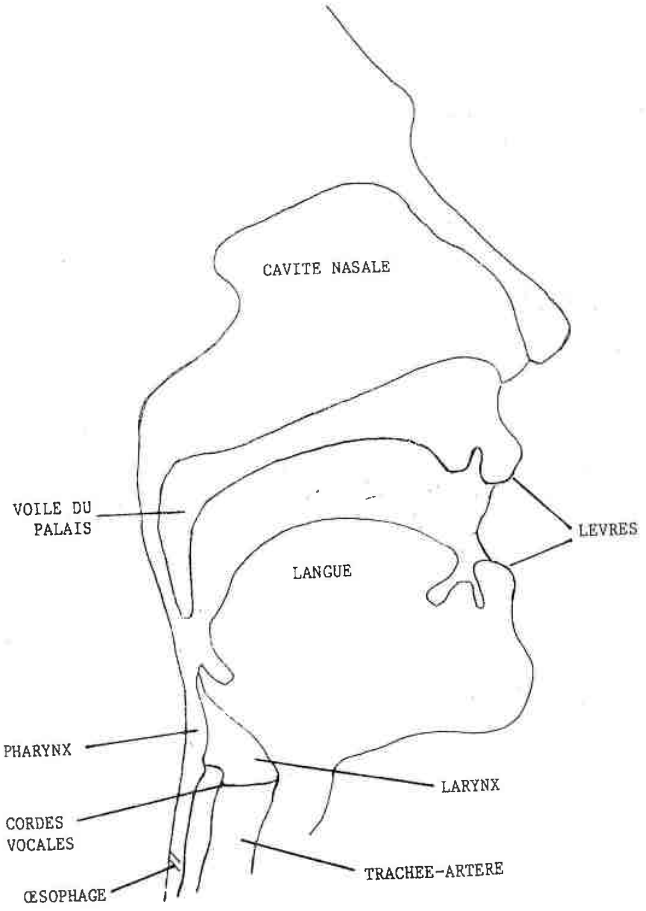


Figure A2: L'appareil vocal d'après [Sundberg 80]

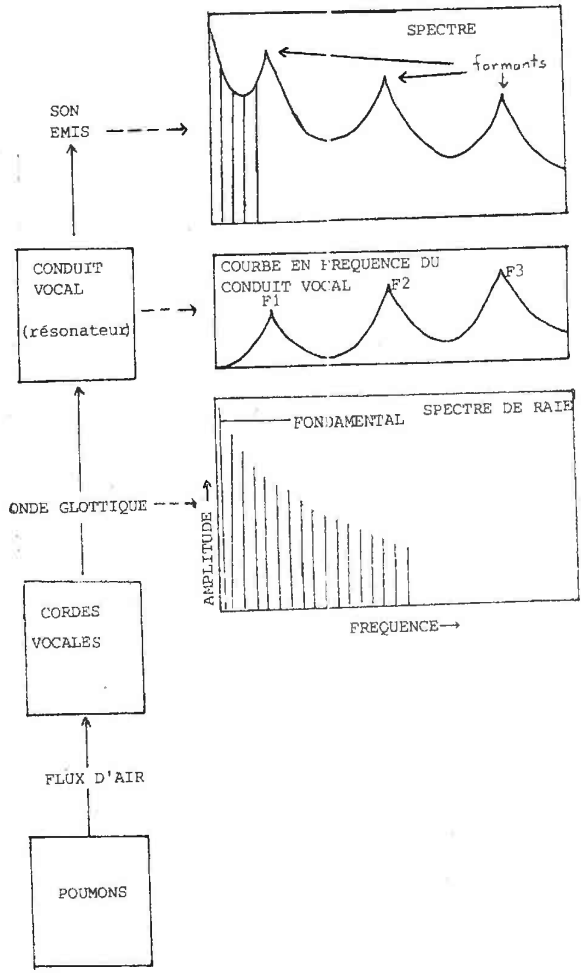


Figure A3: Production d'une voyelle d'après [Sundberg 80]

23. PRODUCTION DES SONS

a) Les voyelles

Les cavités du conduit vocal (pharynx, cavité buccale, cavité nasale) agissent sur l'onde glottique comme un résonateur privilégiant ainsi certaines de ses harmoniques, le spectre de fréquence du son émis montre alors un certain nombre de maxima d'énergie qu'on appelle les formants. (cf. Figure A3, A14 et A17) La grande plasticité des organes vocaux (mâchoire, langue, palais, luette, lèvres...) permet une distinction articulatoire discriminante des voyelles.(cf. Figure A4 et A7). On distingue essentiellement deux critères d'articulation.

- L'aperture (espace compris entre le haut de la langue et le palais)
- Le lieu d'articulation c'est-à-dire la position de la langue : antérieure (vers les dents) ou postérieure (rejetée en arrière vers le voile du palais.

La figure A5 montre la dispersion des voyelles francaises selon ces deux critères.
La figure A6 montre les mêmes voyelles projetées selon la valeur de leurs deux premiers formants (F1 et F2). Le rapprochement de ces deux figures (A5 et A6) illustre bien le lien qui existe entre la formation articulatoire des voyelles et leurs caractéristiques formantiques.
Pour la production des voyelles nasales :
/ $\tilde{\varepsilon}$ /(IN) ; / $\tilde{\text{œ}}$ /(UN) ; / $\tilde{\text{a}}$ /(AN) ; / $\tilde{\text{o}}$ /(ON)
Le voile du palais s'abaisse et permet à l'air de passer simultanément dans les conduits buccal et nasal. Ce couplage a pour effet d'affaiblir les formants à fréquence élevée (anti-formant nasal). L'ouverture du velum, du fait de sa lenteur, introduit alors une certaine instabilité dans le son les positions articulatoires sont proches respectivement des voyelles orales / ε /, / œ /, / a /, / o / entrainent, pour des systèmes de reconnaissance basés sur un codage fréquentiel, des confusions qui paraissent étonnantes à un auditeur humain, par exemple, dans le cadre de notre projet (cf. partie B), entre "CINQ" et "SEPT" ou entre "UN" et "DEUX".

b) Les consonnes

Le flux d'air provenant des poumons peut rencontrer plusieurs types d'obstacles au niveau du conduit vocal entrainant ainsi un jet d'air tourbillonaire. Tandis que les voyelles sont caractérisées acoustiquement poar une onde quasi-périodique, les consonnes au contraire montrent une forte présence de bruits.

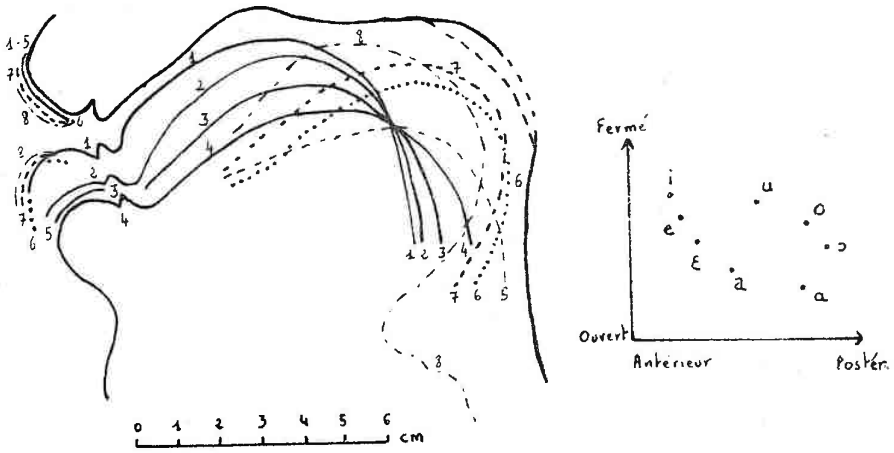


Figure A4: Formes du conduit vocal pour les voyelles orales

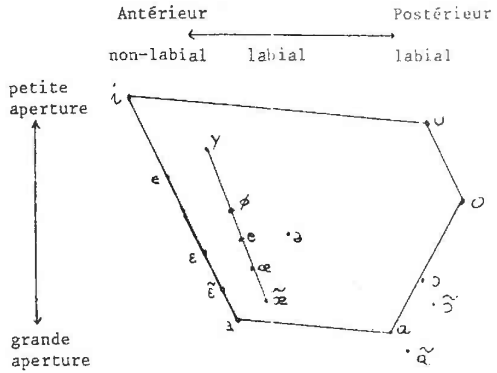


Figure A5: Trapezoide vocalique du français (articulation) d'après [Carton 79]

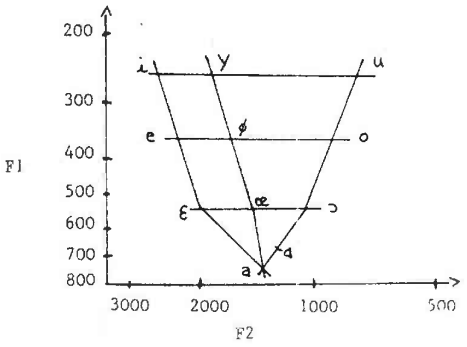


Figure A6: Triangle vocalique du français (formants)

Selon le type d'obstacle et sa durée on peut classer les consonnes françaises en fricatives, plosives, sonnantes.

* Les fricatives sont les consonnes les plus fortement bruitées, elles sont produites par le chuintement qu'émet l'air pulmonaire au passage d'un resserrement important du conduit vocal (cf. Figure A7).
  - au niveau labio-dental (lèvres-dents) pour /f/, /v/
  - au niveau alvéolaire (langue-devant du palais) pour /s/, /z/
  - au niveau prépalatal (langue, dessus du palais) pour /ʃ/ , /ʒ/

Les fricatives peuvent être sonores (/v/, /z/, /ʒ/ ou sourdes /f/, /s/, /ʃ/) suivant qu'elles sont ou non accompagnées d'une vibration des cordes vocales.

* Les plosives sont produites par une fermeture complète mais de brève durée du conduit vocal suivie d'une ouverture brutale. L'obstruction a lieu : (cf. Figure A7)
  - au niveau bi-labial (lèvres) pour /p/, /b/
  - au niveau apico-dental (langue-dents) pour /t/, /d/
  - au niveau vélaire (arrière de la langue-voile du palais) pour /k/, /g/.

Pendant la période occlusive de la plosive l'air peut tout de même s'accumuler en deçà du lieu d'obstruction (dans la bouche par exemple pour un /b/) faisant ainsi vibrer les cordes vocales (buzz), il s'agit alors de plosives sonores :/b/, /d/, /g/. Dans le cas contraire il s'agit de plosives sourdes :/p/, /t/, /k/ leur spectre est alors caractérisé par une période de silence (correspondant à l'occlusion) suivie d'une explosion très courte (le burst). Ceci explique certains problèmes de segmentation aux frontières des mots dans les systèmes de reconnaissance de mots isolés. En effet, lorsqu'un mot est terminé par une plosive le système de détection parole/non parole confond le silence occlusif avec la fin du mot. Cette confusion est particulièrement fréquente pour le vocabulaire utilisé par le système présenté en partie B constitué des chiffres francais, en effet cinq, sept, huit, et quelquefois quatre présentent cette configuration.

* Les sonantes nasales résultent de l'obstruction du conduit buccal et de l'ouverture du velum qui permet l'échappement de l'air par le conduit nasal, elles sont toujours sonores :
  - Obstruction au niveau des lèvres /m/
  - Obstruction au niveau langue-dents /n/

* Les sonantes liquides sont très variables et dépendantes du contexte, elles

VOYELLES

	Antérieures		Postérieures	
	Ecartees	Arrondies	Ecartees	Arrondies
Orales très fermées	i riz	y rue	********	u roue
Orales fermées	e ré	ø peu	********	o pot
Orale moyenne	********	ə le	********	********
Orales ouvertes	ɛ raie	œ peur	********	ɔ port
Orales très ouvertes	a patte		********	ɑ pâte
Nasales	ɛ̃ vin	œ̃ un	ɑ̃ blanc	ɔ̃ bon
Semi-voyelles	j hier	ɥ huit	********	w oui

CONSONNES

	Occlusives			
	Bi-labiales A1	Apico-dentales B2	Palatale D-E 5.6-7	Vélaires E 8-9
Sourdes	p passe	t toux	********	k cou
Sonores	b basse	d doux	********	g goût
Nasales	m masse	n nous	ɲ signe	********

	Fricatives				
	Labio-dentales A2	Alvéolaires C-D 5	Alvéolaire latérale B5	Prépalatale D5	Uvulaire F 9-10
Sourdes	f fer	s assis	********	ʃ chou	********
Sonores	v verre	z Asie	l la	ʒ joue	R rat

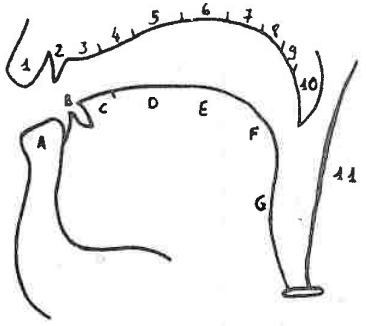


Figure A7: Classification articulatoire du français d'après [Carton 74]

sont généralement sonores.

- /l/ est produit par une obstruction partielle du conduit buccal au niveau langue-palais
- /R / est produit par une vibration de la langue ou du velum suivant les accents régionaux.

24. VARIABILITE

Les processus physiques très divers et complexes mis en oeuvre lors de la production de la parole (cf. Paragraphe 23), rendent le signal vocal extrêmement variable et multi-forme. Cette variabilité se manifeste à plusieurs niveaux :

* Variabilité due aux locuteurs

- Intra-locuteur

Plusieurs critères peuvent faire varier la voix d'un même locuteur, par exemple :

- Le timbre de la voix évolue suivant l'heure de la journée, le degré de fatigue et l'activité du locuteur.
- Il est bien connu qu'une émotion telle que la peur ou l'angoisse agissant directement sur les muscles du larynx, (gorge serrée) affecte le timbre et le rythme de la voix [Williams 70]. Cette propriété est d'ailleurs à la base de plusieurs types de ''détecteurs de mensonges''.
- Une autre altération importante de la voix est due aux maladies qui affectent les organes phonatoires (laryngite) ou le conduit vocal (cavités nasales avec le coryza).

-Inter-locuteurs

La variabilité du signal vocal s'accroit encore lorsqu'on étend son étude à plusieurs locuteurs. Il intervient alors des différences de modes de prononciations selon :

- L'âge,
- Le sexe : de nombreuses études [Lienard 72, Pister 84, Tigziri 85,...] montrent l'influence du sexe du locuteur sur la fréquence du fondamental (en moyenne 100 à 150 Hz pour un homme, 200 à 300 Hz pour une femme) et sur les formants. cf. Figure A8 et A9.
- Les accents régionaux et nationaux : Gupta et Mermelstein ont étudié l'influence de la provenance des locuteurs (canadiens français et anglais) sur les taux de reconnaissance de mots isolés [Gupta 82]. En français, on peut noter également de fortes variations : nasalisation, prononciation des ''e '' muets pour les locuteurs méridionaux, /R/ roulé des bourguignons, etc... cf.[Pister 84, Carton 79]
- Les voix pathologiques : il faut enfin tenir compte de variations extrêmement importantes des caractéristiques vocales dues à certaines anomalies physiques ou neurologiques des locuteurs :

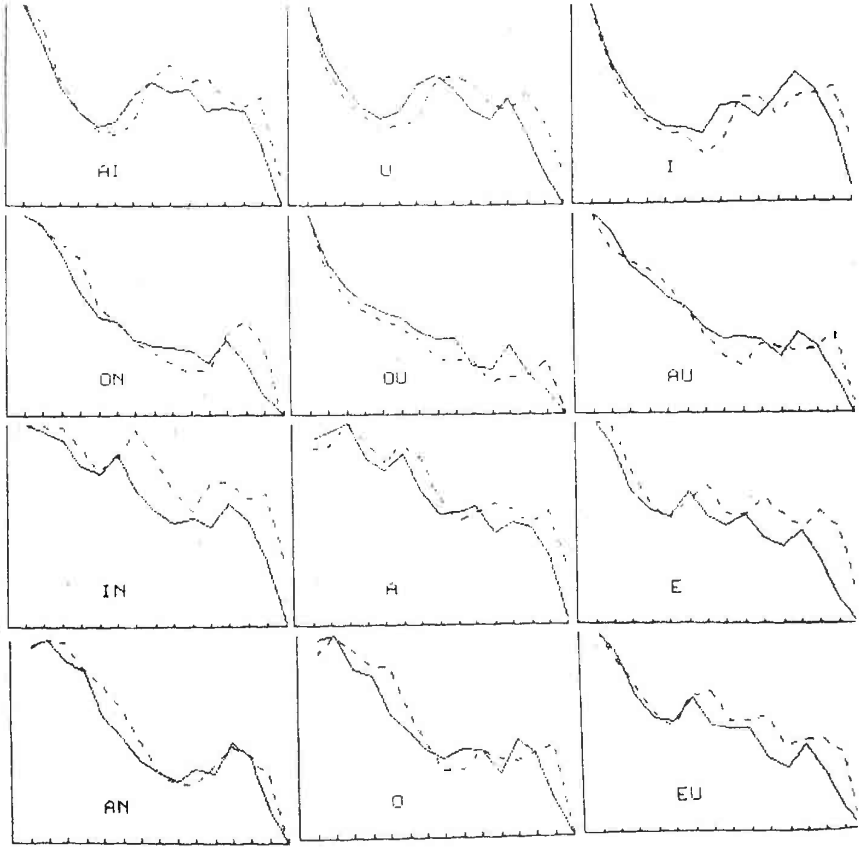


Figure A8: Spectres vocodeur de 9 hommes et 6 femmes [Pister 84]

malformations bucco-faciales : laryngectomies, surdité précoce entraînant des troubles de la phonation...(Signalons que nous développons au CRIN un système interactif de rééducation vocale des enfants non-entendants : SIRENE. cf. [Haton MC 85]).

* Variabilité due au contexte

- Les mots

Lorsqu'une phrase est prononcée de facon naturelle c'est-à-dire en enchaînant les mots il se produit des altérations de leur prononciation aux frontières de ces mots.

- Les liaisons apprises : Pour des raisons d'euphonie, dans le cas de deux mots consécutifs juxtaposant deux voyelles, l'usage veut qu'on prononce la consonne finale du premier mot ; exemples :

(les) (animaux)		≠	(les animaux)
le animo			lezanimo
(vient)	(il)	≠	(vient-il)
vjɛ̃	il		vjɛ̃ til

- Les co-articulations spontanées : en discours continu à débit normal, les positions idéales des organes phonatoires (cf. Paragraphe 23) ne sont pas toujours atteintes ; la tendance à l'économie articulatoire (due aux contraintes mécaniques d'inertie du conduit vocal) et les défauts de synchronisation des mouvements phonatoires entrainent des variations aux frontières des mots par rapport à la prononciation des mêmes mots isolés:

(bec)	(de)	(canard)		≠	(bec de canard)
bɛk	də	kanar			begdəkanar

Elision et nasalisation :

(une)	(heure)	(et)	(demie)		≠	(une heure et demie)
yn	oer	ɛ	dəmi			ynoerɛnmi

Elision et dévoisement :

(pot)	(de)	(camp)	≠	(pot de camp)
po	də	kɑ̃		potkɑ̃

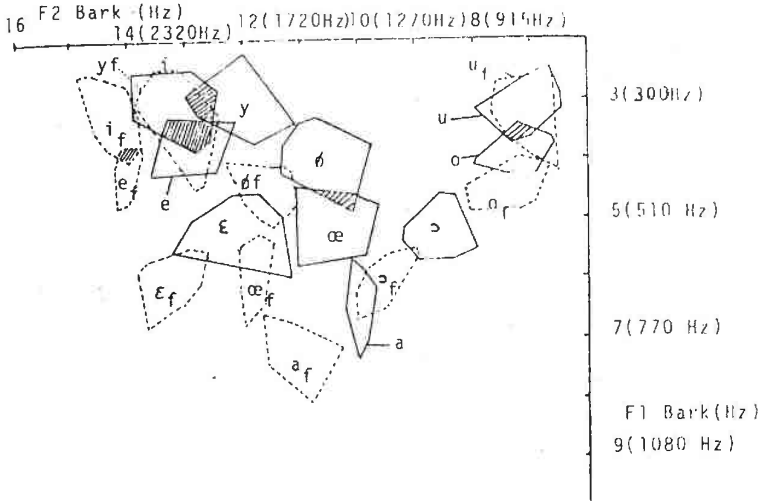


Figure A9: Dispersion des voyelles orales dans le plan F1/F2
(Traits pleins: Hommes ; Traits pointilles : Femmes)

- Les phonèmes

Les mêmes contraintes physiologiques sont à l'oeuvre dans la prononciation des phonèmes contigus à l'intérieur d'un mot. Les altérations phonologiques des voyelles en contexte ont été largement étudiées : Travaux de Peterson et Barney [Peterson 52], de [Stevens 71], [Broad 74], etc pour l'anglais. Citons pour les français : [Lienard 72, Pister 84, Tigziri 85, Fohr 86].

Par exemple dans le mot ''doute '' le deuxième formant du /u/ est beaucoup plus élevé que dans le cas d'un /u/ isolé (cf. Figure A10). L'altération peut aller jusqu'à l'élision partielle ou complète suivant le débit de parole, comme par exemple le /i/ dans ''architecture ''.

C'est dans le cas des consonnes que la variabilité est la plus importante. Certains phonèmes tels que /R/ possèdent presque autant d'allophones que de contextes. Elle a cependant été peu étudiée dans le cadre multilocuteur voir [Pister 84] [Fohr 86]. Voici quelques exemples :

. voisement du /k/ dans ''aqueduc''
akədyk ----> agdyk

. dévoisement du /b/ dans ''absurde '' (apsyrd)

. élision du /k/ dans ''infarctus'' ($\tilde{\varepsilon}$ fartys)

. insertion d'un /s/ dans ''petite '' (pətsit)

L'influence de l'anticipation sur une consonne de la voyelle qui la suit est telle qu'elle peut rendre indiscernable deux fricatives telles que /s/ et /ʃ/ en contexte /ʃi/ et /su/. (voir les spectrogrammes de la figure A11). Cette anticipation peut se faire sentir jusqu'à trois phonèmes à l'avance, par exemple le /s/ sera prononcé différemment dans ''strict'' et dans ''structure ''.

Dans le cadre de notre projet certaines variations multilocuteur sont particulièrement influentes, en particulier aux frontières des mots lors de la procédure d'acquisition ; par exemple certaines personnes, essentiellement des femmes, prononcent des /s/ avec très peu d'énergie haute fréquence au point qu'ils apparaissent comme des silences sur les spectrogrammes et qu'ils ne déclenchent pas l'acquisition en début de mot (''cinq'', ''six'', ''sept''). Chez certains locuteurs également, l'occlusive initiale de ''deux'' et de ''neuf'' est précédée d'un léger ''coup de glotte '' (buzz) qui déclenche l'acquisition plus tôt que pour la plupart des locuteurs.

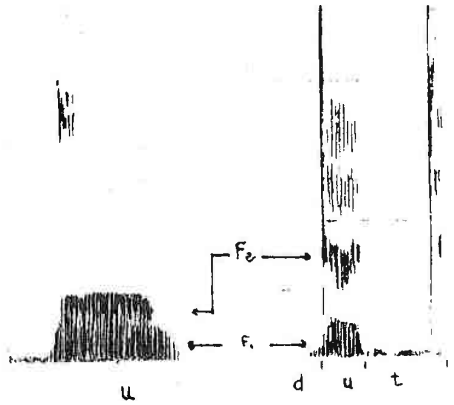


Figure A10 : Spectrogramme de la voyelle /u/ hors-contexte et en contexte dental (/dut/) près de 35% de variation.

[Fohr 86]

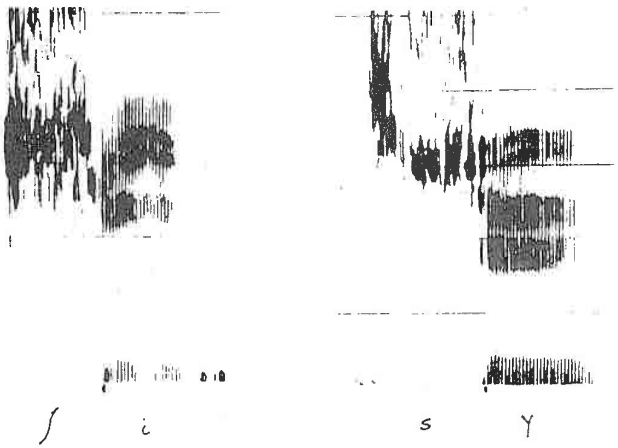


Figure A11: Spectrogrammes des logatomes /ʃi/ et /su/ : On peut remarquer que la limite inférieure du bruit est à la même fréquence dans les deux cas.

[Fohr 86]

25. REDONDANCE

Malgré son importante variabilité, la parole reste très robuste et très peu sensible au bruit grâce à la redondance du signal vocal. Pour une restitution fidèle de la parole, un débit d'informations de 100 000 bits/s est nécessaire. Le codage fréquentiel de ce signal (cf.paragraphe 26) est de débit 10 ou 20 fois moindre, il est cependant parfaitement intelligible comme le prouvent les expériences de codage-resynthèse faites au CRIN par A.Gourinda.

Il faut noter que la redondance de la parole est encore accrue par les niveaux supérieurs de la compréhension (lexical, syntaxique, sémantique) et par la méta-communication parallèle, (gestes, posture, mouvement des lèvres visible,...).

26.CODAGE

Dans les systèmes de reconnaissance de la parole la source d'informations est un capteur (microphone) qui transforme les variations de pressions (les sons) en signal électrique (cf. figure A12). Le codage consiste à extraire de ce signal les informations pertinentes en réduisant la redondance (cf. 25). Il faut noter que cette réduction élimine en grande partie d'autres informations (sur l'idendité, l'état émotionnel du locuteur...) qui ne sont pas utilisées pour la reconnaissance.

Plusieurs types de méthodes sont couramment utilisées dont on trouvera une description détaillée dans [Schafer 74].

* Les méthodes temporelles (mesure des pics, des passages par zéro...) bien que d'extraction rapide sont généralement abandonnées au profit de méthodes plus fiables.
* Les méthodes fréquentielles consistent à extraire du signal, son spectre à court terme (cf. figure A13) révélant les formants caractéristiques des phonèmes (cf paragraphe 22) ; elles ont pour avantage d'être conformes à l'analyse que fait l'oreille humaine et de ne conserver que l'information pertinente, elles ont pour inconvénient de nécessiter des calculs souvent très longs et d'interdire ainsi un traitement de la parole proche du ''temps réel''. Citons comme méthodes d'obtention du spectre à court terme :
  - l'analyse par prédiction linéaire (L.P.C.) est basée sur la modélisation du système phonatoire appliquant à une source d'excitation la fonction de transfert du conduit vocal, le principe de l'analyse L.P.C consiste à trouver les paramètres de cette fonction de transfert. (voir par exemple [Makhoul 72]).
  - L'analyse par transformée rapide de Fourier (F.F.T) permet à partir du signal échantillonné de définir n valeurs du spectre à l'aide de n échantillons x (n) de f(t)

$$X_i(w) = \sum_{n=0}^{N-1} x_i(n)\ w(n)\ e^{-jwn}$$

où w(n) est la fenêtre (la plus utilisée étant la fenêtre de Hamming.
  - L'analyse cepstrale est basée sur l'hypothèse que le signal vocal est issu d'une convolution de la fonction d'excitation avec la réponse impulsionnelle du conduit vocal. L'analyse cepstrale séparant les deux termes de cette convolution fournit une bonne information à la fois sur la fréquence du fondamental et sur les pics formantiques. On peut comparer les résultats de ces trois dernières méthodes fréquentielles sur la figure A14.

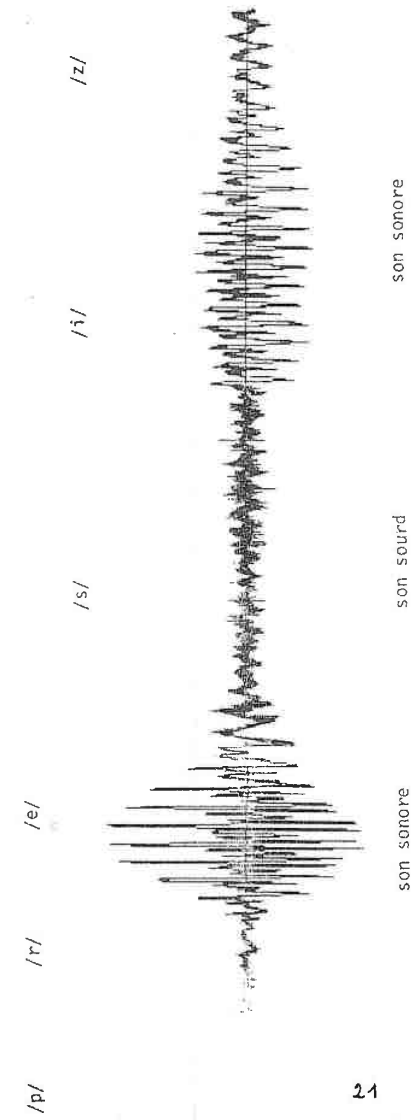
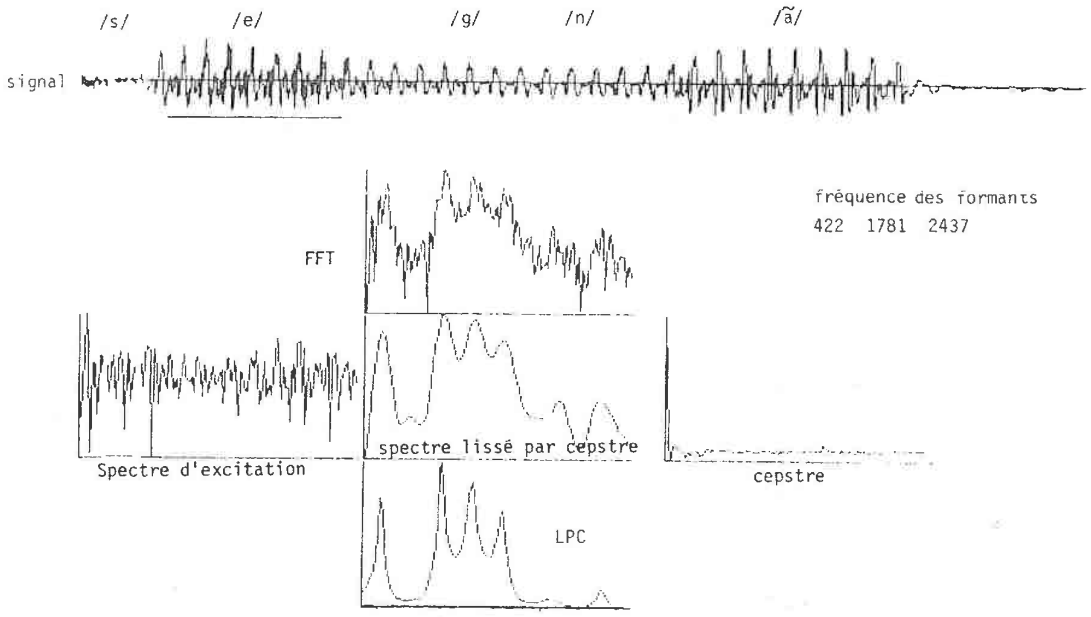
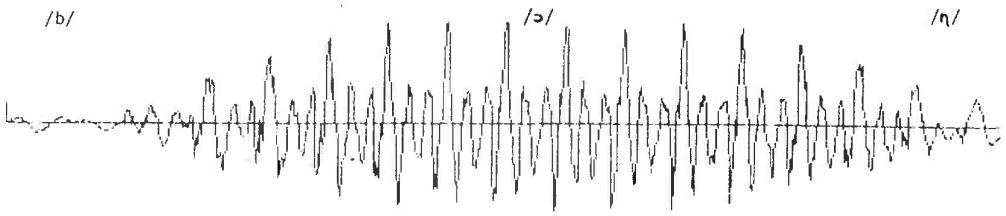


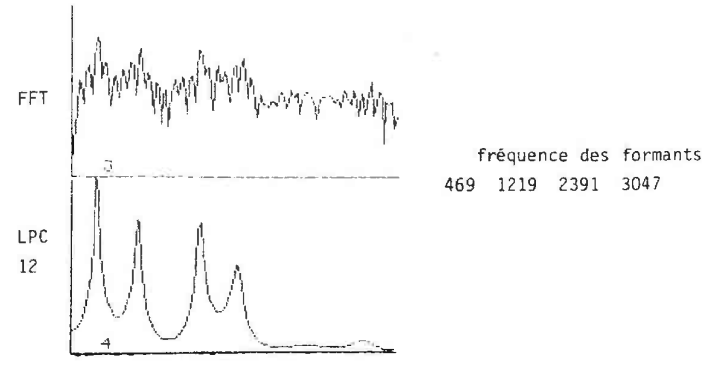
Figure A12 : Le signal de la parole

Figure A 14 : Les résultats des différentes étapes de l'analyse cepstrale sur /e/



23





22

Figure A 13 : LPC à 12 coefficients sur /ɔ/
FFT sur /ɔ/

- L'analyse par bancs de filtres (vocodeur à canaux). Un ensemble de filtres passe-bande convenablement choisis donne une bonne approximation du spectre à court terme, en temps réel. Sa rapidité d'analyse nous a fait choisir cette méthode.
Cependant, le matériel que nous avons utilisé(vocodeur SLE) présente quelques imperfections dont :
    + une finesse insuffisante dans la répartition de l'énergie (16 canaux).
    + un défaut d'accentuation des harmoniques élevées entraînant une mauvaise analyse des hautes fréquences (pour les fricatives féminines, en particulier).

Visualisation des spectres:

Pour la représentation visuelle des mots, des phonèmes, trois dimensions sont importantes :
+ l'énergie du signal
+ l'échelle de fréquence } spectre à court terme
+ le temps

On peut visualiser un mot :
* Par un graphique "en relief" montrant l'évolution du spectre dans le temps (figure A15) ; (peu précis).
* Par un tableau de valeurs numériques représentant la succession des vecteurs de coefficients (Figure A16) ; (peu clair).
* On préfère généralement la représentation sous forme de spectrogrammes en deux dimensions fréquence-temps avec un noircissement plus ou moins important pour représenter les valeurs de l'énergie du signal. Soit grâce à un appareil spécialisé (Figure A17) soit en codant les valeurs numériques par un caractère plus ou moins dense (Figure A18). Avec ces deux méthodes, les formants apparaissent sous la forme de plages sombres. C'est cette dernière méthode que nous avons choisie pour la visualisation des spectrogrammes de notre projet.

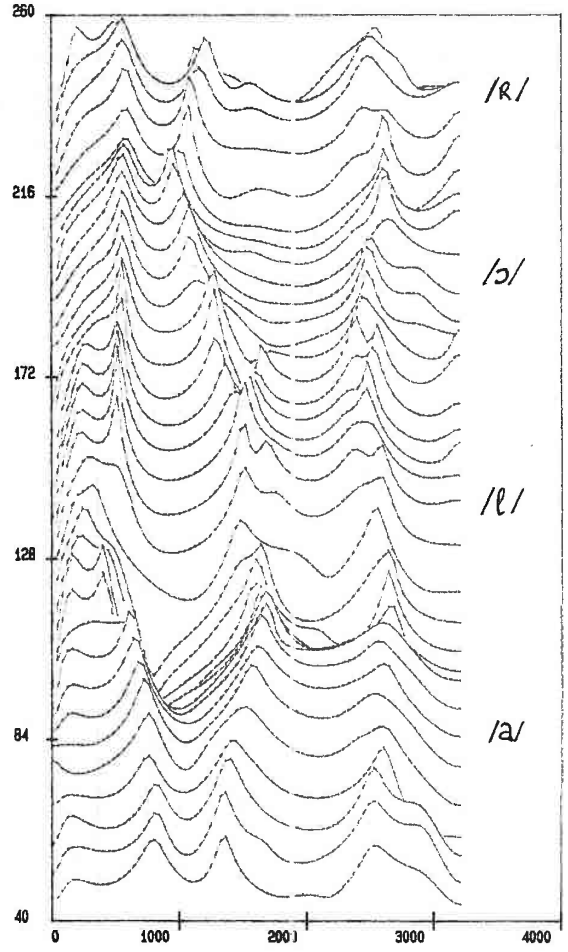


Figure A15: Spectrogramme pseudo-spacial du mot "alors"

Figure A16: Spectrogramme numérique de la phrase: ''Poupée bizarre''

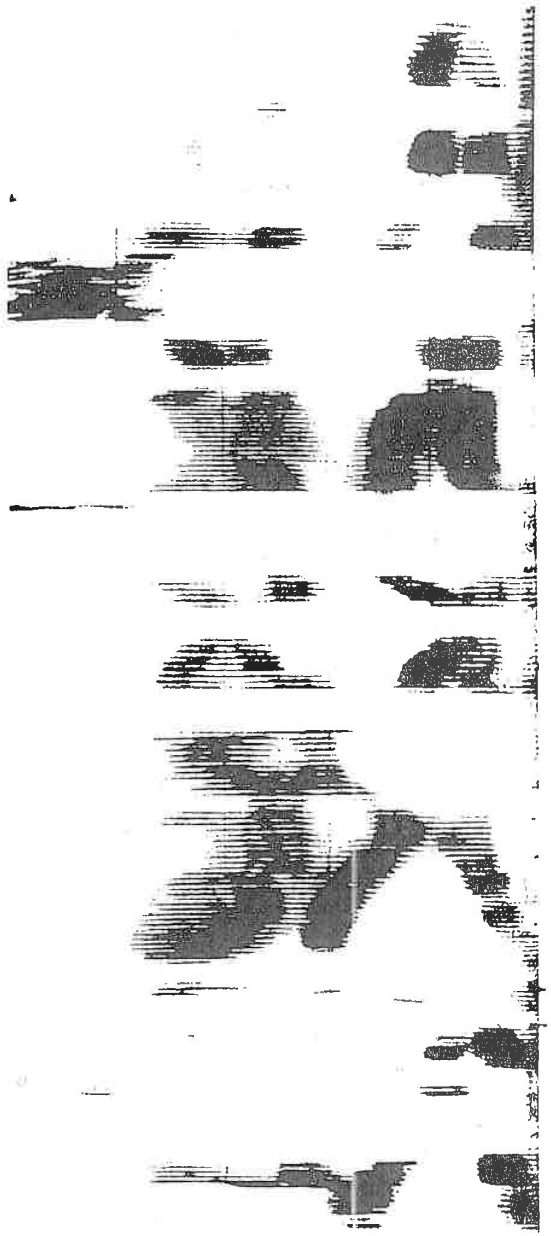


Figure A17 : Spectrogramme de la phrase correspondant à la transcription de :

''Le courier arrive en retard en ce moment''

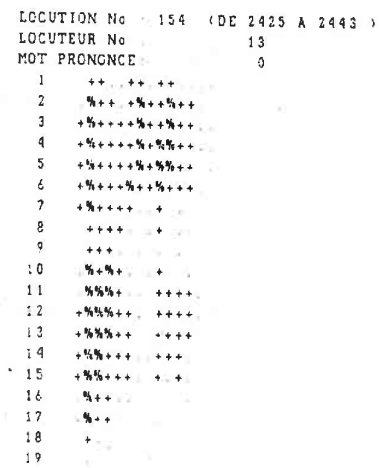


Figure A18: Spectrogramme codé du mot ''zéro''

CHAPITRE 3

LA RECONNAISSANCE DE LA PAROLE

31.GENERALITES

Le concept de communication orale homme-machine se généralise et atteint même le grand public et le secteur industriel comme le montre la figure A19. La reconnaissance de la parole est un des deux aspects de cette communication (l'autre étant la synthèse), le moins bien maîtrisé et celui pour lequel de gros efforts de recherche restent encore à faire.

On peut remarquer avec L. Miclet [Miclet 85] que la reconnaissance de la parole opère sur un ''matériau fait de données réelles, bruitées provenant de capteurs imparfaits;[...] et que, d'autre part ces données sont supposées avoir un sens ,c'est-à-dire qu'il existe des experts sachant les classer, les interpréter, les utiliser''. Malheureusement ces experts en compréhension de la parole que nous sommes tous, sont en grande partie incapables d'expliciter et de formaliser leur expertise.

Il s'agit donc d'automatiser le processus de reconnaissance en transformant les données jusqu'à les réduire à l'information cherchée.

Ce processus,bien que faisant appel à des disciplines différentes (traitement du signal, statistiques, théorie des langages...) s'organise selon un schéma constant (cf figure A20):

La reconnaissance consiste à identifier c'est-à-dire à classer des formes l'aide desquels on identifie éventuellement des formes plus complexes (mots, phrases...). Les informations nécessaires à cette opération sont généralement issues d'une phase préalable d'apprentissage qui consiste à mémoriser une description caractéristique de chaque classe de formes à reconnaître (représentant, règles structurelles, paramètres de reconnaissance etc...). Ces caractéristiques ayant été extraites à partir d'un ensemble de formes dont la répartition en classes est connues (corpus d' apprentissage).

On regroupe sous le terme de système de reconnaissance de la parole un grand nombre de systèmes très différents tant du point de vue de leurs applications que des méthdes qu'elles mettent en oeuvre. Les paragraphes suivants en donnent une classification selon deux axes : Les domaines d'application (paragraphe 32) et les stratégies utilisées (paragraphe 33).

RÉCAPITULATION: MARCHÉ MONDIAL DE LA RECONNAISSANCE ET DE LA SYNTHÈSE AUTOMATIQUE DE LA PAROLE (R.A.P. ET S.A.P.)

(en millions de dollars 1980)

	R.A.P.		S.A.P.		R.A.P. + S.A.P.	
	1985	1990	1985	1990	1985	1990
Automatismes industriels	46	202			46	202
Terminaux informatiques	68	1 103	16	185	84	1 288
Bureautique	28	512	11	308	39	820
Applications télématiques	–	100	50	224	50	324
Automobile	7	85	21	71	28	156
Grand public	214	473	223	323	437	796
Sous-total	363	2 475	321	1 111	684	3 586
Autres (médical, enseignement	36	742	64	222	100	964
Total	399	3 217	385	1 333	784	4 550

Figure A19: Marché mondial du traitement automatique de la parole

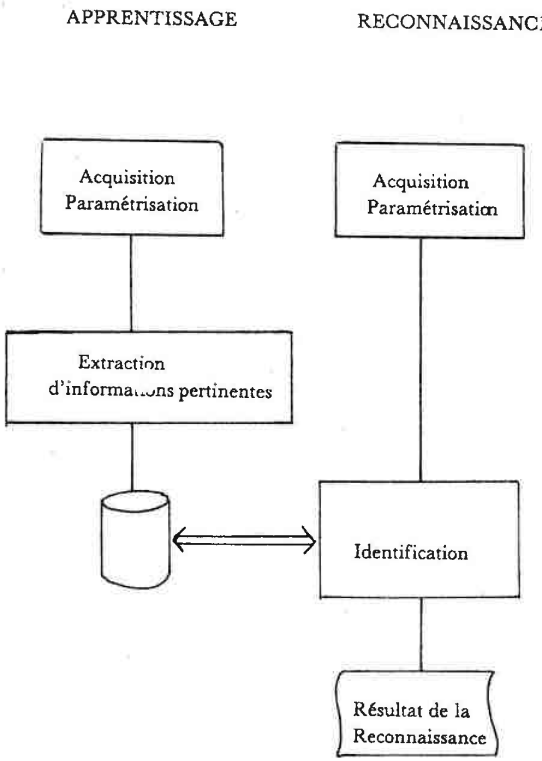


Figure A20: Schéma général de la reconnaissance de la parole

32. LES DOMAINES D'APPLICATION

321. Le langage décrit

Le critère principal de classification des systèmes dont dépendront à la fois les applications et en grande partie les stratégies, est la nature du langage dont on veut rendre compte.

a) Structure du langage

On peut distinguer par ordre de complexité structurelle croissante:

- Les mots-clés: Il s'agit dans ce cas de faire réagir le système à un mot choisi parmi un ensemble limité et connu; ce mot peut être prononcé seul, isolément ou parmi un continuum de parole.

- Les langages formels: Il faut ici, en plus, prendre en compte une structure du langage donc un aspect syntaxique. Un langage formel ne possède qu'un nombre réduit de règles syntaxiques connues a priori (par exemple un langage informatique).

- Le langage naturel : Le nombre et la complexité des règles syntaxiques, le nombre d'exception à ces règles permettant de traiter une langue naturelle humaine est tel qu'il est impossible à un ordinateur de les gérer. On se limite actuellement aux langages pseudo-naturels qui en sont une restriction à un sous-ensemble de vocabulaire et de formes syntaxiques beaucoup plus limité.

b) Continuité du flot de parole

Une des difficultés des systèmes de reconnaissance consiste à isoler les unités phonétiques ou lexicales dans le flot de parole, c'est pourquoi la continuité du langage à reconnaître est une caractéristique importante.

- Mots isolés : C'est évidemment le cas le plus simple, les mots sont prononcés séparément c'est-à-dire avec une légère pause entre chacun d'eux. Cette façon de procéder est très artificielle et contraignante pour le locuteur, mais facilite grandement la tâche de l'ordinateur pour lequel la localisation des mots se ramène à une séparation signal-bruit.

- Mots enchaînés : Le locuteur est dans ce cas autorisé à enchaîner un petit nombre de mots sans pause.

- Discours continu : Dans ce cas, le signal vocal est un semi-continuum dans lequel n'apparaissent pas en général les frontières de mots, ce qui implique pour la segmentation en unités phonétiques le recours à des traitements complexes et à des informations de niveau supérieur (lexical, syntaxique, sémantique, pragmatique).

Applications

- Les systèmes de reconnaissance par mots-cles isolés peuvent être utilisés pour des numéros de code ou pour des commandes simples de machines dans les cas où les mains sont déjà occupées ou pour des handicapés.(cf figure A21). Ils sont assez contraignants pour le locuteur mais fournissent leurs résultats en temps réel, ce sont de plus les seuls systèmes actuels pour lesquels les taux de reconnaissance permettent une application industrielle.

- Les systèmes de reconnaissance de mots enchaînés ont les mêmes applications que les précédents mais permettent une élocution plus naturelle (un code de quelques chiffres peut être prononcé en une seule fois, par exemple).

- La reconnaissance d'un langage formel ou pseudo-naturel en discours continu est la plus proche du comportement humain, c'est donc évidemment la plus riche d'applications.(Dialogue homme-machine,réponses automatiques pour un centre de renseignements, dictée automatique [Charpillet 85], etc...).

322. Le vocabulaire

a) La taille

Par leur influence sur les temps de réponses et sur l'encombrement mémoire, les grands vocabulaires (1000 mots et plus) nécessitent des traitements particuliers.

b) La complexité

La complexité d'un vocabulaire influe grandement sur les taux de reconnaissance . Un vocabulaire, même étendu, peut être simple si ses mots sont bien différents les uns des autres, inversement un vocabulaire

Exemples d'application de la reconnaissance de la parole

Ces applications utilisent des systèmes de reconnaissance de mots isolés, seuls disponibles actuellement sur le marché.

* Commande et contrôle de machines
  - Programmation numérique de machines-outils
  - Montage de paquets (aéroports, centres de tri)
  - Jouets (commande de voitures téleguidées)
  - Commande de microscope
  - Dialogue pilote-avion (projet
  - Commande de fonctions (essuie-glaces, etc...) en automobile

* Inspection, contrôle de qualité
  - Chaînes d'assemblage (automobiles : General Motors, circuits intégrés)
  - Inspection de compresseurs
  - Inspection de tubes de télévision

* Saisie de données
  - Relevé de données en cartographie
  - Entrée d'informations boursières, météorologiques
  - Entrée de séquences de chiffres (numéros d'immatriculation, etc...)
  - Machine à écrire automatique (projet)
  - Gestion de stock

* Contrôle à distance
  - Composeur téléphonique automatique (projet)
  - Commande de téléviseur (projet)
  - Gestion de stock par téléphone pour V.R.P. (projet)
  - Utilisation domestique

* Gestion
  - Entrée vocale sur microordinateur de gestion
  - Bureautique (cf. projet KAYAK de l'INRIA)

* Sécurité
  - Accès en zone réglementée (en association éventuelle avec un système de vérification de locuteur)
  - Applications militaires

* Aide aux handicapés
  - Aide à la lecture labiale
  - Aide à l'apprentissage de la parole chez l'enfant sourd
  - Commande orale de prothèses, de chaises roulantes, etc...

Figure A21: Exemples d'applications de reconnaissance de la parole en mots isolés

[Haton 82]

comprenant beaucoup de mots mono-syllabiques ou de paires minimales (mots ne différant que d'un phonème) est considéré comme complexe. La qualité de l'enregistrement (environnement bruité ou non , locuteurs coopératifs ou non , etc..) peut ajouter encore à la complexité du vocabulaire.

323. Les locuteurs admis

Une des grandes limitation des systèmes actuels tient au nombre de locuteurs qu'ils acceptent.

a) Les systèmes monolocuteurs ne reconnaissent qu'un locuteur, celui qui a effectué l'apprentissage. Leurs applications sont donc réservées aux systèmes mis à la disposition d'une seule personne : par exemple contrôle de fonctions annexes en automobile, commande de chaise roulante pour handicapé ...

b) Les systèmes multilocuteurs parmi lesquels on distingue encore:

Les systèmes plurilocuteurs qui admettent lors de la reconnaissance, les locuteurs ayant participé à l'apprentissage. (Applications dans le cas d'un groupe assez restreint d'utilisateurs connus à l'avance).

Les systèmes indépendants du locuteur qui ne limitent pas l'ensemble des locuteurs admis. Toute personne se présentant au micro doit être reconnue, qu'elle ait ou non participé à la phase d'apprentissage. Ces systèmes trouvent leurs applications pour tous les projets grand-public tels que les centres de renseignements, les appareils à la disposition d'un grand nombre d'utilisateurs(jouets,machines..).

Les tests effectués pour ce travail appartiennent à cette dernière catégorie. Le terme ''multilocuteur'' utilisé sans précision sera d'ailleurs à comprendre comme synonyme de ''indépendant du locuteur''.

33. LES STRATEGIES

On distingue généralement deux grandes classes de stratégies de reconnaissance selon qu'elles portent sur des formes acoustiques considérées commes entières ou composées.

331. Stratégies globales

La forme à identifier est dans ce cas considérée comme un tout insécable. Il s'agit le plus souvent d'un mot ou d'un petit groupe de mots. La reconnaissance consiste alors à placer la forme inconnue dans un espace de représentation dont la signification a été déterminée lors de la phase d 'apprentissage; le plus souvent grâce à un algorithme de classification (''clustering''). On distingue encore parmi elles :

- Les représentations utilisant l'espace des paramètres dans lequel chaque forme est assimilée à un point par ses coordonnées. Les problèmes de l'apprentissage et de la décision sont alors résolus dans le cadre connu de l'espace métrique.

- Les représentations fondées uniquement sur les notions de distances ou de pseudo-distances entre formes. Dans ce cas les propriétes de l'espace de représentation dépendent fortement de la définition de la ''distance ''. Nous verrons en particulier (Partie B;paragraphe 624) les conséquences de l'utilisation de l'indice de programmation dynamique sur l'inégalité triangulaire.

Pour plus de précisions sur ces statégies et leurs limites, on pourra lire par exemple [Simon 84] ou [Fukunaga 73].

332. Stratégies analytiques

Dans ce cas, le message à identifier, mot ou phrase, est considéré comme un ensemble structuré d'unités plus simples elles-mêmes éventuellement encore composées d'unités élémentaires: les primitives. Pour la reconnaissance d'une phrase cette structure pourra être par exemple: Phrase-mot-phonème. Selon la structure à expliquer et les primitives

choisies (phonème,syllabe) les outils peuvent être très variés:

- Analyses syntaxiques: Les formes (le plus souvent des phrases) sont décrites par une grammaire formelle c'est-à-dire un automate.

- Méthodes statistiques et stochastiques : Les algorithmes de classification sont ici encore utilisés par exemple pour l'échantillonnage des primitives. Les modèles markoviens sont également très utilisés pour rendre compte de la structure des mots et des phrases à l'aide d'automates munis de probabilités de transition. (cf paragraphe 443).

- Méthodes cognitives : Les approches récentes utilisent fréquemment les méthodes cognitives issues de la collaboration entre les recherches en intelligence artificielle et en psycholinguistique [Le Ny 80], leur but est de tenter de reproduire le mieux possible le processus humain d'analyse du discours. La compréhension du message passe alors par une analyse simultanée à plusieurs ni aux interdépendants (Phonétique, prosodique, syntaxique, sémantique,..). Notons que cette démarche a été suivie au CRIN par J.M. Pierrel pour les systèmes MYRTILLE [Pierrel 81].

Pour plus de précisions sur les méthodes analytiques et structurelles on pourra lire [Miclet 84],[Jelinek 76],[Fu 74].

34. CONCLUSION

Actuellement, seuls sont bien maîtrisés les systèmes de reconnaissance monolocuteurs de mots isolés et connectés utilisant généralement la comparaison globale de mots. Leur fiabilité et leur temps de réponse pour des vocabulaires limités ont permis leur diffusion sur le marché industriel.

Les domaines de recherche actuels s'orientent vers la reconnaissance de mots sur de grands vocabulaires , la compréhension du discours continu, le décodage acoustico-phonétique, la reconnaissance multilocuteur, certains systèmes combinant parfois deux ou plusieurs de ces domaines.

Le projet MIMULE présenté en partie B, est un système de reconnaissance de mots isolés indépendant du locuteur, fonctionnant sur un petit vocabulaire assez complexe (car bruité et monosyllabique) : Les chiffres français. Il utilise comme stratégie, la comparaison globale de mots par programmation dynamique associée à des techniques statistiques de réduction de données.

CHAPITRE 4

LES SYSTEMES MULTILOCUTEURS

41. INTRODUCTION

Nous avons vu au paragraphe 24 que la variabilité de la parole est une des difficultés essentielles qu'ont à surmonter les systèmes de reconnaissance automatique de la parole. Parmi eux, les systèmes multilocuteurs, sont ceux qui ont à gérer le maximum de sources de variabilité (intra et inter-locuteur).

- Pour un système monolocuteur ou "oligolocuteur", c'est-à-dire admettant un seul ou en tous cas moins d'une dizaine d'utilisateurs, tous connus à l'avance, et dans le cas d'un vocabulaire limité, il est possible de stocker en mémoire les caractéristiques sonores de chaque utilisateur. Cette possibilité facilite grandement la tâche de l'étape de reconnaissance-décision car celle-ci met obligatoirement, à un moment donné, le mot inconnu en comparaison avec son équivalent prononcé par la même personne.

- Nous entendrons ici, par système multilocuteur un système qui puisse fonctionner sans apprentissage particulier avec tous les locuteurs d'une population donnée et qui ne présente pas la possibilité citée plus haut de conserver en mémoire les caractéristiques de chaque élocution, soit parce que le nombre d'utilisateurs est très grand, soit parce que ceux-ci ne sont pas connus au moment de l'apprentissage .

A l'heure actuelle, les systèmes commercialisés pouvant accepter un grand nombre de locuteurs sans détérioration importante des taux de reconnaissance, sont encore assez rares. Citons par exemple,comme précurseur pour l'américain, les systèmes de la firme Verbex aujourd'hui disparue [Baker 81] ou, plus recemment la série V8000 de Votan fonctionnant en mots isolés et enchaînés, le VCS-1000 de VCS fondé sur une reconnaissance phonétique, les cartes VRC d'Interstate qui obtiennent 85 % de reconnaissance exacte ou les circuits WTV de Weitek qui reconnaissent en temps réel 90% de la population avec moins de 10% d'erreur; tous ces systèmes fonctionnent sur de petits vocabulaires (une dizaine de mots). La firme ATT commercialise également un système multiréférence de mots enchainés 'Conversant' fonctionnant sur canal téléphonique [Schinke 86].

Au Japon, la firme NEC commercialise une série de processeurs de reconnaissance vocale (SR-1000) adaptés aux applications de télécommunication en mots isolés (vocabulaire moyen) ou mots connectés (petit vocabulaire) [Kasuya 82].

En France, la société Vecsys commercialise le système Moise [Lienard 82] mis au point par le LIMSI dont une version est multilocuteur: Mozart [Mariani 84]; de même la société X-Com distribue le système Séraphine élaboré au CNET [Gagnoulet 84] ; les laboratoires de la CGE à Marcoussis ont également mis au point un éditeur à entrée vocale (20 mots) implanté sur IBM PC [Flocon 86]. Tous trois sont capables de reconnaître un petit vocabulaire prononcé en mots isolés avec un bon taux de reconnaissance (de 85% à 95% selon les conditions d'apprentissage et d'utilisation).

Comme souvent en reconnaissance des formes, les méthodes et stratégies peuvent être inspirées des processus de perception humains. "Si l'on se réfère à l'être humain, la solution se trouve à la fois dans la mise au point de procédures d'adaptation très évoluées et dans la recherche d'invariants, l'un complétant l'autre " [Haton 85].

Toutefois, en l'état actuel de nos connaissances phonétiques et neuropsychologiques, les méthodes qui se révèlent les plus efficaces reposent essentiellement sur la recherche d'estimateurs robustes des propriétés statistiques du corpus de parole utilisé.

Nous aborderons successivement dans ce chapitre trois grands axes permettant de gérer la variabilité multilocuteur : la recherche d'invariants phonétiques (paragraphe 42), les systèmes adaptatifs (paragraphe 43), les approches statistiques (paragraphe 44). A noter que loin d'être mutuellement exclusifs, ces trois pôles peuvent être complémentaires et sont fréquemment intimement liés dans les réalisations concrètes de systèmes multilocuteurs.

Remarque: Les taux de reconnaissance des systèmes fournis dans ce chapitre sont ceux anoncés par les laboratoires ou firmes qui les ont développés. Il est très difficile de les comparer entre eux, ceux-ci étant extremement dépendants de la nature du vocabulaire testé, de la langue utilisée et des conditions de locution (cf 322b).

42. RECHERCHE D'INVARIANTS PHONETIQUES

La découverte de caractéristiques acoustiques invariantes et spécifiques de chaque phonème serait évidemment une réponse séduisante et efficace au problème de la reconnaissance de la parole multilocuteur.

Malheureusement, cette tâche parait extrêmement ardue; la caractérisation d'un phonème par sa fonction distinctive (ensemble d'attributs acoustiques) [Jakobson 63] se révèle peu réaliste ou en tous cas peu utilisable directement. En effet, la reconnaissance de la parole multilocuteur met en jeu toutes les sources de variabilité du signal (cf paragraphe 24). On constate par exemple sur les figures A9 et A22 combien le recouvrement est important et la discrimination difficile sur un ensemble de locuteurs pourtant limité et sur le sous-ensemble phonétique le mieux différencié: celui des voyelles non nasales.

Citons toutefois deux approches récentes de recherche d'indices acoustiques indépendants du locuteur. L'une basée sur les variations d'énergie dans le spectre-vocodeur des voyelles françaises [Rossi 83] et étendue aux consonnes [Fonsale 84]: Le système Multifon effectue l'étiquetage du signal en classes de phonèmes suivi d'une comparaison dynamique globale sur les mots de références ainsi étiquetés et obtient une reconnaissance exacte dans 86% des cas (tests faits sur 9 mots bien différenciés avec 3 locuteurs) [Fonsale 84b]. Cette même approche a été suivie par Makino et Kido pour un système multilocuteur de reconnaissance de grands vocabulaires et obtient un score de 88.1% pour un vocabulaire de 212 mots japonais testé sur 50 locuteurs [Makino 84].

Les méthodes d'intelligence artificielle permettent actuellement de développer une voie analogue : La recherche de règles invariantes pour le décodage acoustico-phonétique. Les systèmes experts permettent une souplesse de caractérisation des phonèmes qui n'était pas possible par la méthode classique des attributs acoustiques. Ils permettent en effet une représentation déclarative et modulaire bien adaptée à l'acquisition incrémentale de connaissances dans des domaines mal formalisés au départ.

On ne sait malheureusement pas sur quels critères l'homme se fonde pour reconnaître la parole, car l'acquisition de l'expertise, faite dans son jeune âge, est en grande partie inaccessible à l'introspection. C'est pourquoi on passe généralement par l'intermédiaire d'un expert en lecture des spectogrammes qui est, lui, capable d'expliquer son raisonnement et et ses stratégies.

Parmi les nombreux travaux récents sur ce sujet citons [Zue 82], [Johanssen 83], les systèmes SERAC du CNET à Lannion [Gillet 84], SONEX du LIMSI [Memmi 84]. Au CRIN nous travaillons à la mise au point du système APHODEX [Fohr 86] réalisé à partir de l'expertise d'un phonéticien de l'institut de phonétique de Nancy qui obtient, en discours continu, multilocuteur, un pourcentage de reconnaissance bien supérieur aux performances des systèmes automatiques actuels.

De manière générale, les travaux fondés sur le décodage acoustico-phonétique par des techniques d'intelligence artificielle sont surtout orientées vers des systèmes multi-niveau de compréhension de la parole continue et obtiennent des résultats de reconnaissance pour l'instant inférieurs aux systèmes de comparaison globale de mots isolés.

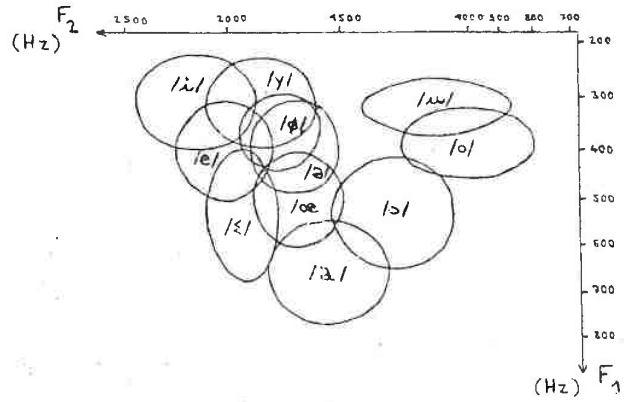


Figure A22: Dispersion des voyelles dans le plan F1/F2 [Malécot 56]

43. LES SYSTEMES ADAPTATIFS

Ce sont des systèmes intermédiaires entre les systèmes monolocuteurs où l'apprentissage est fait en entier pour chaque locuteur, et les systèmes indépendants du locuteur où l'apprentissage est fait une fois pour toutes.

Ils ont pour but de réduire au maximum la phase d'apprentissage imposée à chaque nouveau locuteur en la décomposant en deux parties :

- Une première partie, faite une seule fois, permet de trouver des formes de référence, des paramètres, des règles de transformation, etc... peu dépendants du locuteur ;
- La deuxième partie (phase d'adaptation) est effectuée pour chaque nouveau locuteur et consiste à faire prononcer à celui-ci une phrase connue à l'avance du système. Elle permet ainsi d'ajuster au mieux les formes de référence issues de la première phase, à ce locuteur particulier. Elle est en général assez simple et courte pour ne pas nécessiter la présence d'une personne qualifiée comme c'est le cas pour les apprentissages monolocuteurs.

La phase d'adaptation peut varier de quelques phrases à quelques phonèmes suivant la méthode utilisée. Nous présentons dans ce paragraphe quelques méthodes récentes classées par ordre décroissant de longueur de la phase d'adaptation.

a) Adaptation des formes de référence par cadrage

Il s'agit ici de faire prononcer au nouveau locuteur quelques phrases-tests convenablement choisies. L'apprentissage automatique consiste alors à segmenter et à étiqueter les primitives de chaque phrase-test en alignant des segments de sa description standard avec ceux de la phrase énoncée, à l'aide d'algorithmes de cadrage. Les primitives choisies peuvent représenter des phonèmes [Lecorre 79], [Neel 83] [Andreewsky 85] ou des classes de phonèmes [Wagner 81], [Pister 84] telles que ''plosive voisée '', ''voyelle ''...

La description standard des phrases-tests en primitives doit être faite par un spécialiste qui doit tenir compte des altérations phonologiques possibles et des caractéristiques de la chaîne d'acquisition. Une autre possibilité consiste à engendrer automatiquement la description standard en utilisant la synthèse par règles [Bridle 83], [Lennig 83].

Le problème essentiel des méthodes de cadrage reste l'apprentissage initial du dictionnaire des primitives servant de référence pour la phase

d'adaptation. Ces primitives doivent en effet être peu sensibles au locuteur pour permettre un cadrage efficace.

b) Génération automatique des formes de référence

Cette méthode consiste à engendrer les primitives phonétiques propres au nouveau locuteur à partir d'un sous-ensemble d'entre elles (une seule phrase ou même quelques phonèmes), ceci grâce à un ensemble de règles de transformation appliquées sur les phonèmes prononcés par ce locuteur.

Ces règles de transformation peuvent être obtenues par analyse statistique (regression, corrélation) d'un grand ensemble de phonèmes multilocuteurs, permettant ansi de trouver entre eux des relations indépendantes du locuteur [Furui 80].

Le système SDS d'IBM pour la reconnaissance multilocuteur de grands vocabulaires (5000 mots) choisit par classification les vecteurs servant de base au modèle de Markov acoustique, lors d'une phase d'adaptation au nouveau locuteur de 20 minutes [Hoskins 85].

D'autres règles permettent par exemple de déterminer les deux premiers formants de toutes les voyelles d'un locuteur à partir de seulement trois voyelles prononcées [Gerstman 86], ces règles étant éventuellement différentes pour les hommes et les femmes [Weinstein 75].

Citons enfin des méthodes d'adaptation entièrement transparentes à l'utilisateur:

Partant de l'hypothèse que pour une même voyelle, seule la longueur du conduit vocal varie en fonction des différents locuteurs ,sa forme restant constante, [Wakita 77] propose une normalisation des phonèmes à partir de la longueur du conduit vocal et de la fonction d'aire estimées directement sur le signal de parole à reconnaître.

De même, le système HARPY [Lowerre 77], apprend dynamiquement les paramètres dépendants du locuteur pendant que celui-ci utilise le système. L'utilisateur débute la session avec un vocabulaire de références multilocuteur général qui évolue progressivement par moyennages successifs vers un vocabulaire plus proche du locuteur.

44. LES APPROCHES STATISTIQUES

441. Principe

Les systèmes de reconnaissance multilocuteurs sont généralement basés sur une comparaison entre la forme à reconnaître et un ensemble de formes de référence (prototypes). Leurs performances (temps de réponse, encombrement mémoire) dépendent fortement:

- de la nature de ces prototypes : Vecteurs pour les stratégies analytiques, matrices pour les stratégies globales, graphes pour les méthodes stochastiques...
- de leur nombre.

L'encombrement d'un prototype est fonction des procédures de codage du signal utilisées. Cette première compression de l'information n'est cependant pas suffisante pour faire de la reconnaissance multilocuteur car on ne peut coder toutes les formes du vocabulaire prononcées par tous les locuteurs de la phase d'apprentissage. On a donc recours à des algorithmes de réduction de données qui regroupent les formes en classes homogènes et ne conservent que certaines caractéristiques de chacune d'elles (représentant,frontières,...), permettant de réduire encore la quantité d'information à mémoriser tout en s'affranchissant relativement des variations inter-locuteurs.

Nous abordons dans les paragraphes suivants différentes méthodes statistiques de gestion de la variabilité, classées selon la nature des formes de référence . Le paragraphe 442 présente les méthodes basées sur une classification de vecteurs aboutissant ou non à l'étiquetage de la parole en phonèmes; le paragraphe 443 aborde les méthodes stochastiques où les formes de référence sont des graphes probabilistes; le paragraphe 444 présente les méthodes statistiques globales où la classification porte sur les mots codés sous formes de matrices (spectrogrammes).

Notons une fois de plus que ces méthodes ne sont pas exclusives les unes des autres et qu'on peut même trouver des systèmes de reconnaissance utilisant un même algorithme de classification à deux niveaux différents et portant sur des formes différentes.

442. Classification de vecteurs

La forme élémentaire est ici le vecteur de parole c'est-à-dire le résultat de l'analyse du signal sur une petite fenêtre (quelques millisecondes) au moyen de techniques de codage appropriées (LPC, FFT, filtres ...cf paragraphe 26). On obtient ainsi pour chaque fenêtre, un certain nombre de coefficients (le vecteur) déterminant la dimension de l'espace de représentation.

On peut distinguer deux classes de méthodes selon qu'on recherche l'étiquetage de ces vecteurs en phonèmes ou pseudo-phonèmes (paragraphe 442-a) ou qu'on s'affranchit de cette identification (paragraphe 442-b).

a) Techniques d'identification

Il s'agit, à partir d'un ensemble d'apprentissage de vecteurs de parole étiquetés, de trouver des critères de discrimination assez fiables pour être appliqués aux vecteurs à reconnaître (stratégies analytiques):

- Séparation par hyperplans.

On suppose être ici dans un espace de représentation métrique qu'on désire partitionner en sous-espaces afin de rendre compte des différents types de vecteurs. Pour cela on cherchera à déterminer des hyperplans discriminants qui séparent les classes (cf figure A23). En cas de classes non séparables directement, ce qui est fréquemment le cas, on peut décider soit de choisir l'hyperplan qui minimise les erreurs de classement, soit de chercher des séparatrices linéaires par morceaux [Duda 66] voire quadriques (cf figure A24), ce qui bien entendu, allonge considérablement les temps d'apprentissage et de reconnaissance. [Morii 85] obtient, à l'aide d'une méthode de ce type, 81.4% de reconnaissance pour des classes de phonèmes japonais et les utilise comme base pour une reconnaissance de grand vocabulaire (274 noms de villes).

- Décision bayesienne.

Les formes à reconnaître sont des vecteurs x dans un espace métrique, distribués selon les lois de probabilité P(x/w). Il s'agit de mesurer la probabilité d'appartenance du vecteur x à la classe w donnée par le théorème de Bayes:

$$P(w/x) = P(x/w) * P(w) * 1/P(x)$$

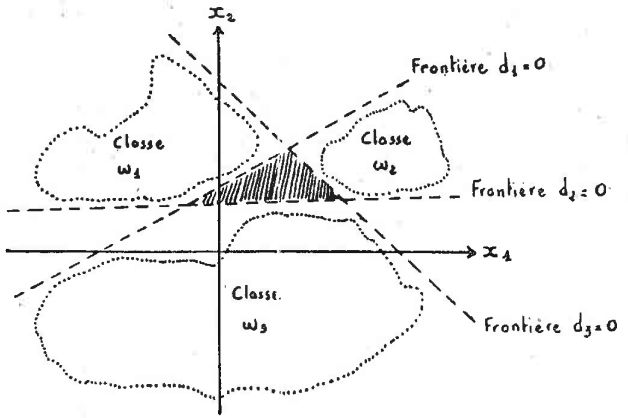


Figure A23: Séparation à l'aide de classificateurs linéaires

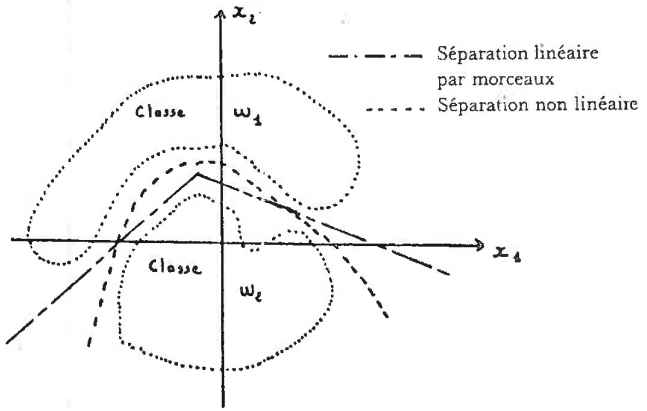


Figure A24: Séparation linéaire par morceaux et non linéaire

Les P(x) sont considérés comme équiprobables, l'estimation des P(w) peut être calculée aisément sur le corpus d'apprentissage. Les probabilités P(x/w) sont généralement supposées être des distributions gaussiennes (cf par exemple le ''Word Based Acoustic Processor'' d'I.B.M. [Bahl 79]).

Cette méthode et la précédente sont fondées sur des hypothèses assez fortes concernant l'espace de représentation (espace métrique, classes peu empiètantes, distributions gaussiennes), la reconnaissance multilocuteur de la parole avec ses capteurs imparfaits, réunit rarement ces conditions, c'est pourquoi on préfère généralement la méthode suivante moins exigente quant à ses a priori.

- K plus proches voisins

Cette méthode consiste à prendre la décision d'appartenance en fonction de l'ensemble d'apprentissage sans faire d'hypothèses a priori sur la forme des classes ou leurs séparatrices.

Le vecteur à reconnaître est assimilé à un point dans l'espace de représentation. On cherche alors autour de lui les K plus proches voisins dans l'ensemble d'apprentissage. Chacun d'eux étant étiqueté par un identificateur de classe, la forme à reconnaître sera décrétée appartenir à la classe la plus fréquente parmi les K ainsi déterminés.

Cette méthode simple et efficace, possède toutefois le défaut d'être en temps de calcul et place mémoire, proportionnelle au nombre de formes d'apprentissage. Celles-ci étant justement nombreuses dans les systèmes multilocuteurs, on effectue normalement une classification sur l'ensemble d'apprentissage, dont on conserve seulement un représentant étiqueté par classe (le prototype). L'efficacité de cette méthode, beaucoup moins coûteuse, dépend des notions de distances et de représentants choisies et de l'algorithme de classification utilisé: [Pister 84] obtient 72% de reconnaissance sur des phonèmes multilocuteurs à l'aide d'un algorithme de classification ascendante hiérarchique portant sur 20 locuteurs. [Lelièvre 81] avec le même type d'algorithme obtient 79% sur les deux premiers phonèmes pour 15 locuteurs. Les algorithmes non-hiérarchiques (cf chapitre 5) sont également très utilisés: [Niles 83] obtient moins de 10% d'erreur sur les classes de phonèmes américains à l'aide de l'algorithme Isodata. [Gupta 78] utilise un algorithme à seuil pour l'identification des voyelles, [Sugamura 83] à l'aide d'un algorithme du même type suivi d'une comparaison dynamique obtient 95 % de reconnaissance en mots isolés sur un vocabulaire de 641 noms de villes (tests faits sur 4 hommes),le projet Sherpa de dictée automatique [Andreewsky 85] reconnait 83% des phonèmes (en 3 propositions) à l'aide d'un algorithme de classification à seuil orientée par les buts et utilisant une métrique appropriée à la phonétique.

Ces méthodes d'identification de phonèmes peuvent être utilisées comme prétraitement afin de ne sélectionner que quelques candidats pour une étape ultérieure plus fine: [Messaoudi 86] procède à une classification (algorithme des K-means) des vecteurs en pseudo-phonèmes qui sont ensuite utilisés dans un modèle de Markov; il obtient ainsi 94.5% de reconnaissance exacte sur un vocabulaire de 51 mots (tests faits sur 6 locuteurs en mots isolés). De même [Lockwood 84] utilise un algorithme de classification en 3 classes phonétiques grossières indépendantes du locuteur afin de limiter le nombre de recherches ultérieures dans le cas de grands vocabulaires.

b) Quantification vectorielle

Utilisée à l'origine pour la transmission de parole à bas débit (jusqu'à 300 bauds), la quantification vectorielle a ensuite été à la base de nombreux systèmes de reconnaissance (principalement multilocuteurs) en raison de sa capacité à réduire l'information manipulée.

Cette méthode est fondée sur la construction d'un ou plusieurs ''code-book'', c'est-à-dire de répertoires de vecteurs de parole obtenus en classifiant l'ensemble ou des sous-ensembles des vecteurs du corpus d'apprentissage. Cette classification, à l'inverse des méthodes précédentes, ne tient pas compte de la ''signification'' de chaque vecteur ; l'identification c'est-à-dire l'étiquetage en phonèmes ou pseudo-phonèmes n'est pas ici recherchée, la quantification vectorielle est donc réservée aux techniques de reconnaissance globales de mots où l'identification est faite au niveau supérieur. Plusieurs méthodes sont utilisées :

•) L 'apprentissage peut consister à faire une classification sur tous les vecteurs de toutes les élocutions d'un même type de mot de vocabulaire. Lors de la reconnaissance, le mot inconnu est codé par chacun des code-book c'est-à-dire que chacun de ses vecteurs est assimilé au vecteur le plus ressemblant du code-book, les distorsions ainsi engendrées sont cumulées. Le mot inconnu est déclaré du type qui a servi à la construction du code-book présentant la plus petite distorsion totale [Burton 83], [Shore 83].

Cette méthode est rapide et économique mais assez sommaire car elle ne prend pas en compte l'information temporelle puisque l'ordre des vecteurs n'intervient pas dans l'évaluation de la distorsion, ainsi deux mots comme ''six'' et ''ici'' peuvent difficilement être distingués. Elles donnent toutefois des résultats intéressants si on l'utilise comme premier étage de sélection de candidats pour des méthodes plus fines: Programmation

dynamique,([Boyer 87] obtient ainsi une première étape de sélection multilocuteur avec 79% de bonne reconnaissance sur les chiffres français) ou, le plus souvent modèles de Markov [Levinson 83], [Tassy 86], ... [Miclet 83] montre par exemple qu'un code-book de 30 vecteurs par mot permet d'obtenir une reconnaissance exacte parmi les trois premiers candidats dans 90% des cas en multilocuteur et dans 95% en plurilocuteur.

*) Afin de pallier l'inconvénient principal de la méthode précédente, [Burton 85] propose de normaliser en longueur les mots de même type, de les découper en N sections égales et d'effectuer une classification de chaque section j de chaque mot de type i, on obtient ainsi un code-book Cij par section de mot, soit N code-books par type de mot. Le mot inconnu est lui-même normalisé, découpé en N sections, chaque section étant codée à l'aide du code-book correspondant, on retient celui qui présente la plus faible distorsion.

Cet algorithme tient compte du temps (ordre des sections) mais pas des variations non linéaires, il est de plus sensible au choix de N et aux décalages dus aux erreurs de déclenchement de l'acquisition par exemple.

*) Une autre méthode consiste à utiliser la quantification vectorielle comme prétraitement, pour diminuer la quantité d'information dans un système basé sur les modèles markoviens (cf paragraphe 443) ou pour accélérer la comparaison dynamique. On crée dans ce cas un code-book général à partir de la classification indifférenciée de tous les mots du corpus d 'apprentissage. Chaque mot du vocabulaire est ensuite codé à l'aide de ce code-book. La matrice triangulaire des distances de tous les vecteurs du code-book entre eux est précalculée. Lors de la reconnaissance, le mot inconnu est codé à l'aide du code-book général puis comparé dynamiquement aux mots de référence; ainsi, moyennant une légère perte d'information, la comparaison est nettement accélérée par le fait que toutes les distances entre vecteurs sont déjà calculées [Pan 85].

Les méthodes de classification utilisées en quantification vectorielle peuvent être fondées sur des algorithmes hiérarchiques le plus souvent dichotomiques descendants tels que l'algorithme de Linde-Buzo-Gray (cf paragraphe 532) [Mari 85] ou sur des algorithmes non hiérarchiques : K-means [Lloyd 82], [Pollard 82], nuées dynamiques [Diday 71],.., si l'on se fixe a priori le nombre de classes à obtenir ; algorithmes à seuils sinon, ceux-ci malgré leurs inconvénients connus (sensibilité à l'ordre des données, chaînage) sont couramment utilisés avec succès [Gupta 78],

[Miclet 84] en raison de leur simplicité, de leur rapidité et de leurs capacités incrémentales. [Tassy 86] obtient ainsi 90% de reconnaissance en 3 propositions sur les chiffres multilocuteurs sans alignement temporel. Cette voie est testée au CRIN dans le cadre de la reconnaissance de la parole continue et fait l'objet d'une thèse à paraître [Gourinda 87].

La plupart de ces algorithmes de classification sont décrits au chapitre 5 avec leurs caractéristiques et spécificités.

443. Les modèles de Markov

La modélisation à l'aide de processus stochastiques est une technique de reconnaissance de mots isolés ou enchaînés qui permet de gérer la variabilité du signal de parole à l'aide d'un automate fini muni de densités de probabilité.

En particulier, le modèle markovien représente la parole par deux suites de variables aléatoires :

$$X(1),X(2),\ldots,X(t) \quad \text{et} \quad Y(1),Y(2),\ldots,Y(t)$$

où on peut dire de façon imagée que les Y(i) sont les observations faites sur le signal et les X(i) sont les positions du conduit vocal qui les ont engendrées.

Un modèle markovien est défini par les données suivantes :

E : Un ensemble de n états ; E={E1,..,En}

A : Une matrice de transition n x n où A(i,j) est la probabilité de transiter de l'état i à l'état j.

B : Une matrice de densité de probabilité n x m ; où m est le nombre de symboles observables et B(i,j) est la probabilité d'observer le symbole j quand le processus se trouve dans l'état i.

Ed,Ef : Deux états particuliers dits initial et final.

En reconnaissance de mots isolés chaque mot de vocabulaire est représenté par un tel modèle, les symboles observés étant les vecteurs de parole le plus souvent issus d'une quantification vectorielle préalable. (cf paragraphe 442)

Le problème de la valuation de l'automate (phase d'apprentissage) est résolu par l'algorithme de Baum [Baum 72] par exemple. Celui-ci détermine les densités de probabilité attachées à chaque état (matrice B) ainsi que les probabilités de transition d'état à état (matrice A) à partir des fréquences d'apparition des symboles sur un vaste corpus d'élocutions du mot associé à cet automate.

Lors de la phase de reconnaissance , le problème est inverse. Etant donné une suite de vecteurs de parole, il s'agit de déterminer la source, et par conséquent le mot associé,qui possède la plus grande probabilité de donner naissance à cette suite d'observations,sachant que la suite des états

pris par le processus est inconnue (d'où le terme de ''Hidden Markov Model'').Le décodeur teste donc chaque modèle M et tente de construire la suite d'états X(1),..,X(t) qui maximise la probabilité de la suite de vecteurs Y(1),..,Y(t). L' algorithme de Viterbi est le plus rapide de ceux qui permettent de résoudre ce problème de façon optimale, le mot reconnu étant celui associé au modèle M qui donne la plus forte probabilité : Y(1),...,Y(t)/M.

De nombreux travaux multilocuteurs récents utilisent les modèles markoviens, le plus souvent couplés à des techniques de quantification vectorielle, tant en mots isolés qu'en mots enchaînés.

Pour les mots isolés, les Bell Laboratories [Rabiner 84], [Juang 85] obtiennent 95.8% de reconnaissance sur les chiffres américains avec une quantification vectorielle basée sur un algorithme de type K-means.

Les mêmes techniques sont utilisées pour les mots enchaînés [Bourlard 84], [Cravero 84], [Jouvet 86]. De même [Mari 85] obtient un taux de 97% sur le vocabulaire des chiffres américains (voix masculines) à l'aide d'un algorithme hiérarchique. Le récent système SDS de IBM utilise un modèle de Markov à 2 niveaux (acoustique et lexical) pour la reconnaissance de grands vocabulaires. Les taux anoncés sont de 92.2% en mots isolés multilocuteurs après une phase d'adaptation de 20 mn (tests faits sur 5 locuteurs avec 5000 mots) [Hoskins 85]. Plus récemment, la même approche permet au système Tangora de reconnaître 20000 mots [Jelinek 87].

On pourra consulter [Tassy 84] pour une bibliographie plus complète sur ce sujet.

Radicalement opposées aux approches cognitives, les méthodes stochastiques présentent l'avantage pour un apprentissage automatique de ne nécessiter aucune connaissance a priori sur les formes concernées exceptée la structure de base de l'automate; on choisit habituellement une structure très générale n'imposant qu'une progression temporelle (cf figure A25). [Levinson 83] a comparé cette méthode (quantification vectorielle plus modèles de Markov) avec une méthode plus classique (classification de mots plus comparaison dynamique), toutes deux en reconnaissance de mots isolés plurilocuteur. Il constate que les taux de reconnaissance sont légèrement en faveur de la seconde (98.5 contre 96.3); au contraire, les considérations d'encombrement mémoire et de temps de calcul influent très nettement en faveur des modèles Markoviens.

Leur principal inconvénient est d'être difficilement interprétable, en effet, les états du modèle n'étant attachés à aucune signification où représentation, l'analyse par un phonéticien des résultats de la reconnaissance est rendue très délicate.

Les modèles connectionnistes, introduits plus récemment et issus des travaux sur le Perceptron, amènent de nouvelles perspectives. Ils sont fondés sur un réseau de cellules cachées simulant le fonctionnement des neurones et s'auto-organisant au fur et à mesure des apprentissages. Ces systèmes, en cours d'expérimentation pour la plupart, aboutissent toutefois à des résultats encourageants [Bourlard 87].

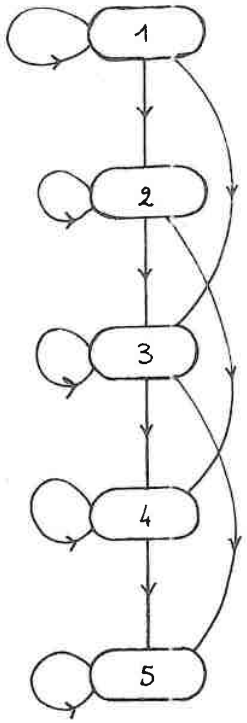


Figure A25: Exemple de modèle de Markov à 5 états

444. Classification de mots

C'est la méthode de reconnaissance de mots isolés la plus ancienne et une de celle qui donne les meilleurs résultats: Voir [Levinson 83] déjà cité pour une comparaison avec les modèles de Markov. [Arcella 86] aboutit aux même conclusions avec un corpus multilocuteur des chiffres italiens: Une légère supériorité (98.3 contre 97.8) des méthodes de classification de mots pourvu que le nombre de références par type de mots soit suffisamment grand (à partir de 7).

Généralisée à partir des systèmes monolocuteurs, cette méthode est fondée sur une comparaison dynamique globale du mot inconnu avec les mots du vocabulaire de référence. Un mot est représenté en mémoire par une matrice de coefficients (n x l) où n représente la taille d'un vecteur de parole et l la longueur du mot en nombre de vecteurs dépendant de la fréquence d'échantillonnage (généralement 100 Hz),(cf partie B,paragraphe 43). On voit que la place mémoire et donc aussi le temps nécessaire à la comparaison sont importants et qu'il est impossible dans un cadre multilocuteur de conserver la totalité des mots prononcés par tous les locuteurs du corpus d'apprentissage. La gestion de la variabilité passe donc comme pour les vecteurs par une phase de réduction des données.

Ce problème de la réduction de l'encombrement mémoire et/ou du temps de calcul à la reconnaissance fait l'objet de nombreux efforts de recherche actuels tant par le développement d'architectures adaptées [Weste 83], [Frison 83], [Feldman 84] que par la mise au point d'algorithmes rapides de recherche du plus proche voisin et d'efficacité équivalente. Parmi ces derniers , deux types de stratégies peuvent être envisagés pour réduire le nombre de comparaisons.

- Structurer le vocabulaire afin d'éviter une comparaison séquentielle exhaustive.

- Réduire la taille du vocabulaire de références en conservant seulement les formes caractéristiques.

a) Structuration

Le principe de ces méthodes consiste à structurer l'ensemble des références de sorte que les propriétés des distances permettent d'éliminer une fraction importante des candidats à la reconnaissance.

Malheureusement,les objets ici représentés sont des matrices de taille variable et les méthodes qui s'appliquent aisément sur un corpus de

vecteurs donc dans un espace métrique de dimension bien définie, sont ici hasardeuses. Toutefois, les algorithmes de comparaison dynamique s'avèrent conférer à cet ensemble des propriétés généralement assez proches de celles des distances où seule l'inégalité triangulaire est contredite et dans une faible proportion (cf pseudo-distance : partie B, paragraphe 62). C'est sur cette constatation que sont fondés les algorithmes hiérarchiques.

Ils consistent à donner à l'ensemble des élocutions du corpus d'apprentissage, une structure d'arbre étiqueté en chaque noeud. L'algorithme général descendant est le suivant : (cf algorithme ascendant au chapitre 5)

- E est l'ensemble des élocutions de départ

- E est partitionné en k classes E1,...,Ek à l'aide d'un algorithme de séparation; de chaque classe Ei est extrait un représentant Xi qui forme un nouveau noeud de l'arborescence.

- Chaque classe Ei est elle-même ensuite partitionnée en k sous-classes et ainsi de suite récursivement jusqu'à obtention de classes toutes réduites à un élément.

On obtient ainsi un arbre d'ordre k (le plus souvent binaire) où chaque noeud est étiqueté par un représentant.

Lors de la reconnaissance, cette structure permet une recherche rapide du plus proche voisin. Cette recherche se fait du tronc vers les feuilles, guidée en chaque noeud par le représentant de la branche qui en est issue. Les algorithmes de complétion d'une classification hiérarchique [Jambu 78] sont ici inutilisables en raison du nombre de comparaisons à effectuer en chaque noeud qui rendrait cette méthode plus lourde que la comparaison séquentielle exhaustive.

On peut par contre utiliser un algorithme avec choix exclusif qui donne une solution approchée mais rapide [Miclet 83], [Lockwood 85] ou un algorithme avec retour arrière (''branch and bound'') [Fukunaga 75], [Kamgar-Parsi 85] qui est optimal mais seulement hypothèse faite d'un espace métrique; [Lockwood 86] montre qu'avec un tel algorithme muni d'une règle d'arrêt on peut diviser par 6 en moyenne le nombre de comparaisons sans déterioration importante du score (4.1% d'erreur au lieu de 3.3%; tests faits avec 10 locuteurs).

Ces algorithmes sont en O(k.logk(N)) où N est le nombre de références à comparer, ils sont évidemment très avantageux par rapport aux algorithmes de comparaison séquentielle exhaustive qui sont en O(N). Cette voie a été testée pour le système MIMULE, (cf algorithme en partie B, paragraphe 63).

Une autre approche possible pour la recherche rapide du plus proche voisin dans l'espace engendré par les comparaisons dynamiques globales entre mots isolés est l'algorithme de recherche par approximation-élimination (AESA) [Vidal-Ruiz 85]. Partant de la constatation que les indices de dissemblance entre les mots sont presque des distances (le corpus testé ne présente pas de violation de l'inégalité triangulaire), l'espace de représentation est considéré comme quasi-métrique. Une inégalité triangulaire ''lâché' est alors utilisée pour éliminer à chaque étape les prototypes ne répondant pas à certains critères de distances minimum et maximum par rapport aux prototypes déjà testés. L'auteur montre que, dans le cas du corpus testé (270 locutions), l'économie en nombre de comparaisons est de l'ordre de 70% sans déterioration du taux de reconnaissance.

Avec les mêmes constatations (moins de 5% de violation de l'inégalité triangulaire sur le corpus d'apprentissage), [Gupta 84] procède à un réétiquetage de chaque mot de ce corpus par le type de ses k plus proches voisins. Il obtient ainsi un score de 94% sur les chiffres américains (avec k=7) en utilisant pour la reconnaissance, un algorithme de recherche rapide du plus proche voisin.

b) Réduction

Le but de ces méthodes est de réduire à l'apprentissage la taille du dictionnaire de références tout en conservant au maximum l'information contenue dans la totalité du corpus. La reconnaissance consiste alors en une comparaison séquentielle du mot inconnu avec chaque mot du vocabulaire conservé, la décision étant généralement prise à l'aide du critère des K plus proches voisins .

La réduction de données effectuée à l'apprentissage utilise les méthodes habituelles de classification, elles consistent à répartir les locutions acquises en classes (ou ''clusters'') homogènes et à conserver seulement un ou quelques prototypes représentant la classe lors de la reconnaissance. Le taux de réduction des données dépend donc de la taille des classes et des algorithmes choisis (il peut aller jusqu'à un pour cent).

Notons ici que tous les algorithmes de classification ne sont pas utilisables, en effet les objets à classer ne sont pas des points connus par leurs coordonnées dans un espace métrique mais des mots-matrices dont seules sont connues les positions relatives en terme de ''pseudo-distance ''. Ces caractéristiques interdisent donc certaines méthodes (comme par exemple la séparation par hyperplans) au profit exclusif de celles basées sur la matrice triangulaire des ''distances'' deux à deux de toutes les locutions entre elles, la notion de distance utilisée étant le plus souvent le score de comparaison

dynamique.

L'ensemble des locutions à classer en une fois peut faire l'objet de plusieurs stratégies : La plupart des systèmes appliquent la classification par type de mot, c'est-à-dire qu'on extrait du corpus total le sous-ensemble de toutes les élocutions du même mot de vocabulaire; pour la reconnaissance de chiffres par exemple, on classera séparément les ''ZERO'',les ''UN'',etc ; Inversement, une méthode globale s'appliquera sur le corpus entier tous types confondus. L'avantage de la classification par type tient à ses performances en terme de temps de calcul et d'encombrement mémoire; si T est le nombre de types et N le nombre de locutions par type de mot, le nombre de comparaisons dynamiques à calculer et donc la taille de la matrice des distances est de l'ordre de $O(T.N^2)$ alors qu'il est en $O(T.N)^2$ pour la classification globale. Cette caractéristique fait que la méthode par type s'avère généralement la seule envisageable étant donnée la taille des corpus d'apprentissage multilocuteurs.

La méthode globale au contraire, possède l'avantage d'une information sur les objets à classer (leur type) d'où une efficacité plus grande dans le cas de classes empiètantes. Dans le système MIMULE, la faible taille du corpus d'apprentissage autorise les deux méthodes, la méthode globale s'avérant aboutir à une classification qui gère mieux les recouvrements de types différents, et à des taux de reconnaissance améliorés (cf partie B, paragraphe 632). Notons au passage que tous les algorithmes ne se prêtent pas à cette prise en compte d'une information a priori sur les objets à classer.

Nous avons vu que la méthode de classification utilisée dépend de plusieurs critères : les propriétés de l'espace de représentation, le nombre d'objets à classer, la connaissance ou non d'information a priori sur les objets; citons encore : la connaissance a priori du nombre de références à obtenir (afin par exemple de fixer une borne supérieure au temps de reconnaissance), la possibilité ou non d'étendre le corpus sans remettre en cause la totalité de la classification (pour les systèmes adaptatifs par exemple), etc...

Il existe deux grandes familles de méthodes de classification : hiérarchiques et non-hierarchiques (ou nodales).

Les méthodes nodales sont les plus utilisées en reconnaissance de la parole en raison de leur simplicité de mise en oeuvre et de leur moindre coût en temps de calcul. Les méthodes hiérarchiques sont très utilisées dans les autres sciences et sont généralement recommandées pour des ensembles à classer de ''petite dimension''. ([Jambu 78] les préconise jusqu'à 10000 objets dans le nuage initial). Elles présentent en effet l'avantage de ne nécessiter aucune connaissance a priori sur la structure du nuage,

contrairement aux méthodes nodales qui fixent a priori le nombre de classes, ou le seuil d'agrégation maximum.

Parmi les méthodes non-hiérarchiques les plus fréquemment utilisées en reconnaissance de la parole, citons la méthode des K-means,(cf description paragraphe 523) [Ball 65], [Levinson 79], [Rabiner 79a]; [Levinson 83] obtient ainsi 98.5% de reconnaissance exacte sur les chiffres en plurilocuteur. Un algorithme récent MKM (modified K-Means) proposé par [Wilpon 85] permet de s'affranchir du choix a priori du nombre de classes, le score obtenu en multilocuteur sur le corpus des chiffres via le téléphone est de 84.2% (soit 1% de mieux que le classique UWA); la même méthode avoisine avoisine 100% en enregistrement de laboratoire et en plurilocuteur. Les nuées dynamiques sont une généralisation de l'algorithme des k-means utilisant des centres multiples pour représenter les classes [Diday 70].

Les algorithmes à seuil (cf paragraphe 521) sont très utilisés également en reconnaissance de la parole. L'algorithme UWA (cf paragraphe 522) [Rabiner 79b],[Wilpon 82] en est une variante dirigeant la recherche du centroide vers les zones de plus forte densité [Rabiner 79b] obtient 80% de reconnaissance avec 10 références par type de mot. [Briant 84] obtient avec la même méthode pour le système Syril, un score de 88% sur les chiffres français plus quelques mots de service. Gupta a amélioré cet algorithme en s'assurant une classification optimale grâce à l'algorithme du lien complet et en remplaçant la notion de centroide par un algorithme à seuil variable et mieux adapté à la recherche des points d'accumulation [Gupta 82]. Le système MIMULE (cf [Divoux 82]) procède également à une amélioration de ce genre en utilisant le centroide MINISOM et un ''rayon d'influence '' adapté à la taille des classes (cf partie B, paragraphe 46). [Mokeddem 85] montre de plus que, quel que soit l'algorithme utilisé (AEC est fondé sur la méthode des K-means et ASC sur un algorithme à seuil) les scores de reconnaissance peuvent être encore améliorés par optimisation de la classification selon une fonction critère bien choisie. Il obtient ainsi un taux de 7% supérieur à la méthode UWA (84% au lieu de 77%) sur le même corpus (les chiffres français).

Les méthodes de classification hiérarchique, bien que plus rarement utilisées en reconnaissance de la parole, permettent également une partition des locutions en classes homogènes, il faut pour cela construire une hiérarchie de partition sur l'ensemble à classer (cf algorithme CAH paragraphe 5) et segmenter l'arbre obtenu, chaque branche formant ulors ne classe. Les critères de segmentation sont très variés, [Lelievre 81] obtient de bons résultats (84% en 1ère position, 89% en 2ème) à l'aide de la partition formée au niveau significatif dont le nombre de classes est le plus proche de six. (Il s'agit d'une classification par type). Les tests portent sur un petit vocabulaire (les chiffres plus 9 mots de service) et sur une

dizaine de locuteurs.

Pour le système Moise [Mariani 84] a testé deux méthodes de classification, l'une nodale (UWA), l'autre hiérarchique (CAH); les deux donnent des résultats comparables: 95% sur le vocabulaire des chiffres plus 7 mots de service en conservant 8 à 12 références par type de mot; la reconnaissance est faite à l'aide de l'algorithme des 2 plus proches voisins, sur 20 locuteurs différents des 100 ayant créés les références. La comparaison avec un algorithme à seuil dans les mêmes conditions montre que ce dernier, plus souple d'emploi et moins coûteux en temps de calcul nécessite cependant plus de références pour les mêmes performances.

Pour le système MIMULE nous avons testés la CAH avec plusieurs critères de segmentation (cf partie B,paragraphe 635). Certains d'entre eux (nombre de points maximum par classe, nombre de classes totales) permettent d'exercer un contrôle sur la classification, d'autres au contraire, (taux de recouvrement minimum) ne sont guidés que par l'information a priori sur les types des feuilles et ne nécessitent aucune décision préalable de l'utilisateur. Ces derniers ne sont applicables que dans le cas de la classification globale tous types confondus.

Notons que la segmentation d'un arbre hiérarchique représente une perte importante de l'information, on passe en effet d'une hiérarchie de partitions à une partition particulière. C'est cette constatation qui justifie les méthodes de structuration (cf 444a) dont la méthode CSA testée pour le système MIMULE (cf partie B, paragraphe 67).

CHAPITRE 5

LES METHODES STATISTIQUES DE CLASSIFICATION AUTOMATIQUE

51. INTRODUCTION

Une des difficultés majeures de la reconnaissance des formes tient à la quantité d'informations brutes issues des capteurs et à la complexité exponentielle de leur combinaison. Le codage et le prétraitement constitue une première étape, trés importante, de filtrage de l'information de ''bas niveau''; celle-ci n'est généralement pas suffisante pour la majeure partie des applications courantes, on peut alors avoir recours à des techniques statistiques de réduction de données appliquées aux informations de ''haut niveau'' (phonèmes, mots,...).

Nous nous limiterons ici à l'étude de la famille de techniques la plus utilisée : Les méthodes de classification. Remarquons tout de suite que la classification ne réduit l'information que dans la mesure où elle est suivie d'une étape de représentation au cours de laquelle chaque classe est ''résumée '' par une faible quantité d'informations (prototype, hyperplan,...). Du point de vue mathématique,une classification sur un ensemble fini I est soit une partition de I (méthodes nodales,cf paragraphe 52), soit plus généralement une hiérarchie de classes emboitées (méthodes hiérarchiques, cf paragraphe 53).

Rappelons également la distinction entre classification et classement : La classification est l'édification de la partition (ou de la hiérarchie de partitions), c'est en reconnaissance des formes le but de l'apprentissage. Le classement est l'opération qui consiste à affecter un élément à un système de classes déjà conçu, c'est le but de la reconnaissance.

Le choix d'une méthode de classification dépend de nombreux critères propres à l'application.

- Les propriétés de l'espace de représentation:

Si les données sont connues par leurs coordonnées dans un espace vectoriel on peut utiliser les méthodes algébriques (cf paragraphe 44) sinon on utilisera les méthodes basées sur une matrice de proximités, précaution prise d'un espace adéquat (dépendant des propriétés de la notion de proximité choisie).

Notons que le premier cas peut toujours se ramener au second moyennant une distance euclidienne par exemple; l'inverse n'étant pas nécessairement vrai.

- Les connaissance a priori sur les données:

Le plus souvent il s'agit du nombre de classes final qui est fixé à l'avance, soit par contrainte sur le temps de calcul ou d'espace mémoire, soit parce que le classifieur possède une connaissance préalable sur la

structure du nuage. Il peut aussi s'agir d'autres critères concernant chaque classe comme par exemple le nombre de points maximum ou la distorsion maximum.

On peut même connaître a priori la répartition complète en classes, le but de la classification est alors de fixer les paramètres qui en rendent le mieux compte. Ce type de connaissance est utilisé dans MIMULE (cf partie B, paragraphe 635) et dans les méthodes d'inférence du lien (cf par exemple [Moreau 84]).

Nous présentons ici quelques méthodes de classification fréquemment utilisées en reconnaissance de la parole et qui ont été évoquées dans les chapitres précédents, au paragraphe 52 nous aborderons les méthodes nodales et au paragraphe 53 les méthodes hiérarchiques.

52. LES METHODES NODALES

Ces méthodes sont en général plus simples et plus rapides que les méthodes hiérarchiques mais nécessitent une connaissance minimum sur les données à classer, de plus, le choix de la notion de représentant influe sur la classification elle-même. Nous décrivons ici deux grands types de classification non hiérarchique, l'algorithme à seuil (paragraphe 521) avec une de ses variantes (paragraphe 522), et l'algorithme classique des K-means (paragraphe 523).

521. L'algorithme à seuil

Cet algorithme utilise un seuil S fixé a priori et une distance d entre les points à classer.

* Le premier point du corpus: X1 constitue la première la première classe C1 dont il est le représentant R1.

$$C1 = \{X1\} \; ; \; R1 = X1$$

* Ensuite pour tout nouveau point X

- Si il existe un représentant Rj de la classe Cj tel que :

$$\text{et} \begin{cases} d(X,Rj) < S \\ \forall k;\ d(X,Rj) < d(X,Rk) \end{cases}$$

Alors Affecter le point X à la classe Cj et recalculer le représentant Rj de cette classe.

- Sinon Créer une nouvelle classe Cp ayant pour représentant X:

$$Cp = \{X\} \; ; \; Rp = X$$

Cet algorithme a pour avantage d'être simple à mettre en oeuvre, rapide, peu encombrant en mémoire, de plus il est incrémental ce qui permet l'adaptation à de nouvelles données sans remettre en question toute la classification.

Il présente par contre les inconvénients suivants:
Il est très sensible à la valeur du seuil qui n'est pas toujours évident à fixer et éventuellement à l'ordre de prise en compte des données. Le représentant étant recalculé à chaque étape, il présente de plus le danger d'effet de chaîne particulièrement gênant pour un classement basé sur le critère du plus proche voisin.

522. Algorithme UWA

L'algorithme UWA (Unsupervised Without Averaging) [Rabiner 79b] est une variante de l'algorithme à seuil où tous les points sont connus (par leurs distances) dès le début.

1. Détermination du centre MINIMAX du corpus E --› R
(voir définition du minimax :Partie B, paragraphe 642)
2. Une première ébauche de cluster (c'est à dire de classe) est obtenue avec tous les points de E situés à l'intérieur de la sphère de rayon S et de centre R.
$C = \{ X \in E / d(X,R) \leq S \}$
3. Le centre MINIMAX de ce cluster C est calculé--› R
4. Les étapes 2 et 3 sont réitérées jusqu'à obtention de la stabilité (Ci = Ci-1). Le cluster Ci est alors définitif et il est retiré de l'ensemble E.
5. Les étapes 1,2,3 et 4 sont réitérées jusqu'à ce que l'ensemble E soit vide.

L'amélioration consiste ici en une recherche de stabilité des classes dans les zones denses pour chaque cluster. De plus, les classes étant remises en question à chaque itération, l'effet de chaîne ne se présente plus. Cependant, le temps de calcul et l'encombrement sont nettement accrus. Cet algorithme présente en outre l'inconvénient de privilégier le ''centre'' du nuage et de sur-segmenter la ''périphérie''. Nous avons tenté de remédier en partie à cet inconvénient grâce à la notion de centre MINISOM (cf partie B,paragraphe 643).

523. Algorithme des K-means

Cet algorithme [Ball 65], [Mac Queen 67],..part de l'ensemble complet des points à classer et d'une distance entre chacun d'eux.

1. Initialisation: On choisit, parmi la population à classer ou non, K représentants (ou directement la partition P1).
2. Chaque point du corpus est ensuite associé au représentant le plus proche, on obtient ainsi une partition P en K sous-ensembles.
3. On effectue ensuite le test d'arrêt qui peut porter sur le nombre d'itérations, la minimisation de la distorsion ou la convergence. s'il est positif, on possède les K classes et leurs représentants,
4. S'il est négatif, on recalcule les K nouveaux représentants de chaque sous-ensemble de la partition et on réitère à partir de l'étape 2.

Ces méthodes permettent un contrôle direct du nombre de classes à obtenir. Leur principal inconvénient tient au choix de départ; puisqu'elles ne sont que localement optimales, le nombre de représentants et leur choix conditionnent toute la classification. Les nuées dynamiques [Diday 70] présentent une amélioration de la méthode des K-means où une classe est représentée par un groupe de centroides et la stabilité améliorée par la notion de ''formes fortes''.

53. LES METHODES HIERARCHIQUES

531. Généralités

Le but de la classification hiérarchique est de représenter un ensemble fini d'objets au moyen de parties de cet ensemble, hiérarchiquement emboitées. Cette représentation correspond d'assez près aux méthodes de taxinomie utilisées dans de nombreuses disciplines scientifiques.

Un ensemble I (par exemple les plantes à fleurs) est divisé en un nombre fini de classes 2 à 2 disjointes (les familles botaniques: Ombellifères, scrofulariacées,...) chaque classe étant elle-même divisée en un nombre fini de sous-classes également toutes disjointes (la famille des scrofulariacées par exemple rassemble de nombreux genres parmi lesquels digitale, véronique ... et mimule) et ainsi de suite jusqu'aux éléments terminaux de la classification (en botanique: l'espèce ou même la variété). Chaque classe est incluse dans une classe et une seule du niveau supérieur; ainsi, du tronc jusqu'à un élément particulier, il n'existe qu'un chemin qu'il s'agit de retrouver lors du classement d'un élément inconnu.

D'un point de vue plus mathématique une hiérarchie peut se définir ainsi:
Soit E un ensemble fini et H un ensemble de parties de E (les paliers), on dit que H est une hiérarchie sur E si les conditions suivantes sont toutes vérifiées:

- $E \in H$ (le palier le plus haut contient toute la population)
- $\forall e \in E , \{e\} \in H$ (les points terminaux appartiennent à la hiérarchie)
- $\forall h,h' \in H$ on a $(h \cap h' \neq \emptyset) \Longrightarrow h \subset h'$ ou $h' \subset h$
(les parties d'une hiérarchie sont soit disjointes soit emboitées)

Exemple: Soit E={X1,X2,X3,X4,X5}

et H = { {X1},{X2},{X3},{X4},{X5},{X1,X2},{X4,X5},{X1,X2,X3},E}

on peut vérifier que H est une hiérarchie sur E (cf figure A26).

Hiérarchie binaire:

C'est une hiérarchie dans laquelle chaque palier est formé du regroupement de 2 des éléments du palier précédent:
$\forall B \in H , \ card(B) > 1 \Longrightarrow \exists (B1,B2) \in HxH$

tel que $B1 \cap B2 = \emptyset$ et $B1 \cup B2 = B$
La hiérarchie de la figure A26 est donc binaire.

Hiérarchie indicée:

On appelle hiérarchie indicée un couple (H,f) où H est une hiérarchie et f une application de H dans $\mathbb{R}+$ telle que:

- $f(h) = 0 \Longleftrightarrow card(h) = 1$
- $\forall h,h' \in H ; \ h \subset h' \Longrightarrow f(h) < f(h')$

On associe donc à chaque palier h un indice numérique f(h) proportionnel à la "distance " entre les deux sous-arbres qu'il relie (voir l'extension de la notion de distance à des classes de points au paragraphe 532). Dans les schémas d'arbres hiérarchiques que nous donnerons, l'indice sera représenté par la hauteur du lien, ainsi la figure A27 représente la hiérarchie H de la figure précédente où on voit que X1 et X2 sont agglomérés plus tôt (indice 1), donc sont plus proches l'un de l'autre que X4 et X5 (indice 1.5).

Définition algorithmique:

D'un point de vue plus algorithmique, une hiérarchie peut également se définir comme une suite (ordonnée dans le temps) de partitions de l'ensemble à classer E. Ainsi, la hiérarchie des figures A26 et A27 pourra être décrite, de façon redondante mais plus conforme au processus de construction, par la suite des partitions qu'elle définit à chaque palier; par exemple :
(P0, P1, P2, P3, P4) avec
P0={{X1},{X2},...,{X5}}
P1={{X1,X2},{X3},{X4},{X5}}
P2={{X1,X2},{X3},{X4,X5}}
P3={{X1,X2,X3},{X4,X5}}
P4={X1,X2,X3,X4,X5}

Une telle suite contient nécessairement à ses extrémités, d'une part l'ensemble total E (ici P4), et d'autre part la partition $P*$ composée de tous les singletons de E (ici P0). Si l'algorithme de construction évolue de $P*$ à E par agglomération des sous-ensembles, la classification est dite ascendante; si au contraire, il évolue de E vers $P*$ par divisions (ou "éclatements") successifs des ensembles, la classification est dite descendante. Ces méthodes bien qu'assez lourdes sont fréquemment utilisées en statistiques en raison de leur efficacité. Elles présentent en effet l'avantage de contenir plus d'informations que les méthodes nodales et pour les CAH d'être indépendantes de la notion de représentant de classe.

Si le but recherché est seulement un ensemble de classes homogènes, la phase de construction de l'arbre doit être suivie d'une phase de segmentation qui consiste à sélectionner le niveau d'agglomération le plus significatif compte tenu de l'application et à considérer comme une classe chaque sous-arbre ainsi segmenté. Différentes stratégies de segmentation faisant intervenir ou non l'information sur les feuilles sont envisagées en partie B, paragraphe 635.

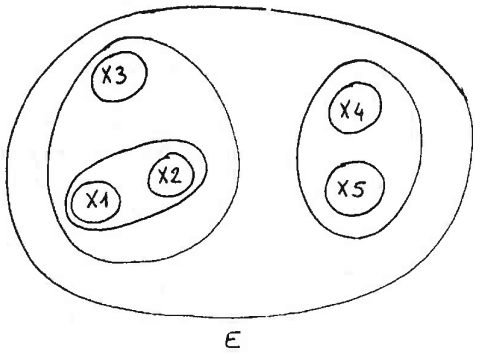


Figure A26: Hiérarchie H sous forme d'ensembles

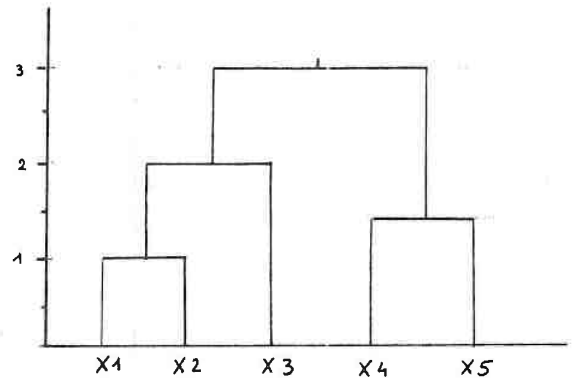


Figure A27: Hiérarchie H sous forme d'arbre binaire

532. Algorithme de Linde-Buzo-Gray

Cet algorithme ([Linde 80]) procède par séparations dichotomiques successives de l'ensemble initial vers une partition de 2^N classes où N est un paramètre du système.

- Initialisation
  On dispose au départ d'un seul ensemble C
  Fixer N
- Répéter N fois
  - Calculer le centroide de chaque ensemble C ---> X
  - Perturber légèrement le centroide X
    $X+\varepsilon$ ---> X1
    $X-\varepsilon$ ---> X2
  - Jusqu'à stabilité de la répartition répéter :
    - Répartir l'ensemble C en 2 sous-ensembles C1 et C2 tels que
      $\forall X \in C1 \quad d(X,X1) \leq d(X,X2)$
      $\forall Y \in C2 \quad d(Y,X2) \leq d(Y,X1)$
    - Calculer les centroides X1' et X2' de C1 et C2
  - L'ensemble C est décomposé en 2 sous-ensembles C1 et C2 qui vont eux-même être décomposés en 2, jusqu'à obtention des 2^N classes.

La construction descendante de cet algorithme permet l'arrêt dès l'obtention du nombre de classes voulu, de plus dans le cas d'une recherche hiérarchique, la reconnaissance emprunte le même sens que l'apprentissage. Toutefois il nécessite de fixer a priori le nombre de classes et ne s'applique pas à toutes les structures de données (ajouter ou retrancher un epsilon au centroide suppose un espace euclidien). C'est pourquoi cette méthode est utilisée pour la quantification vectorielle plus que pour la classification de mots.

533. Classification ascendante hiérarchique

L'édification ascendante d'un système de classes hiérarchisées se fait à partir des eléments terminaux en formant d'abord de petites classes ne regroupant que des éléments très semblables, puis à partir de celles-ci des classes plus lourdes et nécessairement moins homogènes.

A l'etape 0 il y a n éléments terminaux à classifier, à la dernière étape (la n-1 ème) un seul ensemble de n éléments. Chaque étape procède au regroupement des deux classes du niveau précédent, les plus proches (cf algorithme partie B paragraphe 633). On voit que cet algorithme nécessite d'étendre la notion de distance de points à celle de classes de points.

De plus, à chaque étape, 2 classes seulement sont agrégées; il est donc inutile de recalculer la nouvelle matrice de distances à partir des distances initiales entre locutions, des formules de récurrence permettant de calculer la matrice de l'étape i en fonction de celle de l'étape i-1. Cette fonction vérifie lors de l'agrégation des classes Ca et Cb en une classe Cab (= Ca U Cb):

$\forall j,k \neq ab,a,b$

$$D_i(Ck,Cj) = D_{i-1}(Ck,Cj) \qquad (1)$$

$\forall j \neq a,b$

$$D_i(Cab,Cj) = f(D_{i-1}(Ca,Cj),D_{i-1}(Cb,Cj)) \qquad (2)$$

Il existe plusieurs définitions possibles de la distance entre 2 classes de points qu'on trouvera dans [Jambu 78] ainsi que les fonctions f de récurrence correspondantes. Nous nous limiterons à donner ici les quatre formules (2) testées pour le système MIMULE (cf partie B paragraphe 634).

Soient deux classes de points C1 et C2 :

C1= { X1,X2,...Xn }

C2= { Y1,Y2,...Yp }

où les Xi et Yi sont des points de l'ensemble initial à classer.
d(X,Y) représente la notion de distance initiale entre 2 points X et Y. (dans le cadre de notre projet il s'agit du score de comparaison dynamique entre deux locutions).

D(C1,C2) représente la nouvelle notion de distance étendue à deux classes C1 et C2.
A noter que dans le cas de singletons la distance D est calquée sur la distance initiale d.

C1 = {X1} ; C2 = {X2} ==> D(C1,C2) = d(X1,X2)

Rq : Cab désigne la classe formée par l'agrégation de Ca et Cb.

a) Ultra-métrique inférieure (ou saut minimum)

Définition:

$D(C1,C2) = \min_{i,j} (d(Xi,Xj))$

La distance de deux classes est assimilée à la distance de leurs deux points respectifs les plus proches.

Formule de récurrence:

$\forall j \neq a,b$

$D_i(Cab,Cj) = \min_{i,j} (D_{i-1}(Ca,Cj) , D_{i-1}(Cb,Cj))$

b) Ultra-métrique supérieure (ou diamètre maximum)

Définition:

$D(C1,C2) = \max_{i,j} (d(Xi,Xj))$

La distance de deux classes est assimilée à la distance de leurs deux points les plus éloignés.

Formule de récurrence :

$\forall j \neq a,b$

$D_i(Cab,Cj) = \max_{i,j} (D_{i-1}(Ca,Cj) , D_{i-1}(Cb,Cj))$

c) Distance moyenne

Définition :

$D(C1,C2) = (\sum_{i,j} d(Xi,Xj)) / (\text{Card } C1 \cdot \text{Card } C2)$

La distance de deux classes est calculée comme la moyenne de toutes les distances des points de C1 à ceux de C2.

Formule de récurrence :

$\forall j \neq a,b$

$$D_i(Cab,Cj) = \frac{[\text{Card Ca}.D_{i-1}(Ca,Cj) + \text{Card Cb}.D_{i-1}(Cb,Cj)]}{[\text{Card Ca} + \text{Card Cb}]}$$

d) Variance de la réunion des deux classes :

Définition :

D (C1,C2) = Var (C1 U C2)

La distance de deux classes est ici mesurée par la dispersion des points dans la classe formée par leur réunion.

Formule de récurrence :

$\forall j \neq a,b$

$D_i(Cab,Cj) = (A-B) / (Ma+Mb+Mj).(Ma+Mb)$

avec

$A = (Ma+Mj).D_{i-1}(Ca,Cj) + (Mb+Mj).D_{i-1}(Cb,Cj) + (Ma+Mb).D_{i-1}(Ca,Cb)$

$B = Ma.n(a) + Mb.n(b) + Mj.n(j)$

où
n(j) est l'indice de niveau de la classe Cj
Mj est la masse affectée à la classe Cj

PARTIE B

LE SYSTEME MIMULE

CHAPITRE 1

INTRODUCTION

MIMULE est un système de reconnaissance de Mots Isolés MUlti-Locuteur utilisant des méthodes statistiques de classification.

Notre travail a eu pour but de réaliser plusieurs variantes autour d'une structure de base commune (comparaison globale, mots isolés, classification statistique) et de tester leur efficacité respective dans le cadre d'une application précise.

Le contexte de cette application et ses caractéristiques fonctionnelles et matérielles sont présentées au chapitre 2. Afin de ne pas recourir à de fréquentes répétitions nous avons choisi de présenter les différentes variantes de la façon suivante:

Après une description schématique rapide de l'architecture du projet (chapitre 3), nous présenterons dans un premier temps au chapitre 4, la structure commune à toutes les variantes.

Dans un deuxième temps, après une présentation des caractéristiques déterminantes des locutions du corpus d'apprentissage (chapitre 5), le chapitre 6 détaillera les différents choix de variantes ainsi que leur influence sur les taux de reconnaissance.

Le chapitre 7 discute les taux de reconnaissance et la portée des conclusions à la lumière d'autres résultats obtenus avec des corpus de données plus importants.

CHAPITRE 2

LE CADRE DU PROJET MIMULE

21: Contexte de l'application

Il s'agit d'un projet de réponse vocale pour un service de renseignements automatiques. Nous avons limité notre étude aux applications où un dialogue fortement guidé par l'ordinateur permet de n'avoir à reconnaitre qu'un petit nombre de mots-clé prononcés isolément.

Nous avons choisi comme application possible de simuler un centre de renseignements bancaires téléphonés. Le système est chargé d'identifier le client grâce à un code chiffré et d'accomplir un certain nombre de tâches telles que :

- Communiquer les messages en attente qui le concernent,
- Communiquer le solde de son compte courant,
- Prendre en compte une demande de chèquier,
- etc....

Le choix de ces options par le client peut se faire soit par reconnaissance d'un mot-clé avec confirmation, soit par sélection séquentielle (réponse par oui ou non après la proposition de chaque service par l'ordinateur), (cf. organigramme en annexe).

Il est évident qu'un tel "dialogue " se révèle assez fastidieux à l'emploi, surtout pour les locuteurs non habitués à parler dans un microphone et ceci d'autant plus que les performances actuelles de reconnaissance obligent à une confirmation systématique des chiffres reconnus. Cette contrainte limite de façon draconnienne l'extension à des vocabulaires plus grands. Nous ne présenterons ici que la reconnaissance du code secret à 4 chiffres (cf. figure B1) :

Pour chaque chiffre prononcé, le module de reconnaissance propose successivement les 2 réponses les plus probables. Le module de dialogue demande alors confirmation au locuteur de l'une ou des deux.

Si les deux propositions sont rejetées par le locuteur, on lui demande de reprononcer son mot ; si, à nouveau, il rejette les deux propositions qui lui sont faites, la procédure est interrompue et abandonnée.

C'est pourquoi lors des statistiques fournies au chapitre 6, nous donnons toujours les résultats de reconnaissance en deux pourcentages : les performances obtenues en une seule proposition et celles obtenues avec les deux meilleures propositions.

On ne demande évidemment pas confirmation pour les réponses de type oui-non...! Cette confiance est heureusement justifiée par un taux de reconnaissance de 100% pour ce sous-vocabulaire.

- Pour i de 1 a 4 Répéter

  - Répéter

    - Synthèse(''Prononcer le i ème chiffre de votre code '')
    - Acquisition
    - reconnaissance-chiffre ---> X1,X2
    - Synthèse(''Le i ème chiffre de votre code est-il X1 ?'')
    - Acquisition
    - Reconnaissance-oui-non ---> Y
    - Si Y =''oui'' alors Code (i) = X1
    - Si Y =''non'' alors

      - Synthèse(''Le i ème chiffre est-il alors X2 ?'')
      - Acquisition
      - Reconnaissance-oui-non ---> Y
      - Si Y = ''oui'' alors Code (i) = X2
      - Si Y = ''non'' alors Synthèse(''Désolé, je n'ai pas compris'')

  jusqu'à Y = ''oui'' ou nb-de-répétitions = 3
  - Si nb-de répétitions = 3 alors
    - Synthèse(message d'excuse)
    - fin du dialogue

- Synthèse(''Votre code est Code(1),Code(2),Code(3),Code(4)'')
- Suite du dialogue

Figure B1: Algorithme du dialogue

22 : Caractéristiques principales

Le système tel qu'il a été défini, ainsi que ses applications, possède un certain nombre de caractéristiques spécifiques importantes :

- Utilisation potentielle par un grand nombre de locuteurs inconnus a priori. Le système devra donc gérer la grande variabilité (inter et intra-locuteur) des différentes voix et l'impossibilité de posséder un vocabulaire par utilisateur;
- Vocabulaire de petite taille (ici 12 types de mots : les 10 chiffres, oui, non) ;
- Mots prononcés isolément en réponse à une question fermée de l'ordinateur, donc pas d'aspect syntaxique ni de problème de segmentation dans un flot continu de parole mais seulement une segmentation parole-bruit ambiant.

23. Architecture matérielle

Ayant coincidé avec un renouvellement de l'équipement informatique du laboratoire, le déroulement du projet comprend deux ''époques'' :

1) Le premier système : MULOT, a été implanté sur mini-ordinateur MITRA 125 avec une mémoire centrale de 32 K-mots de 16 bits, programmé en FORTRAN sur cartes perforées. L'organe de prétraitement du signal vocal est un VOCODEUR SLE à 16 filtres numériques. Celui-ci fournit tous les centièmes de seconde l'énergie du signal vocal dans les 16 bandes de fréquence décrites sur la figure B2.

2) Le deuxième système : MIMULE, a été implanté sur mini-ordinateur MOTOROLA EXORMACS (micro-processeur 68000) avec une mémoire centrale de 1 méga-octet et programmé en PASCAL. Pour les tests effectués sur les corpus les plus importants (C44 et C60), l'acquisition, la paramétrisation et les comparaisons ont été faites sur Masscomp et programmés en C; les algorithmes de classification et sélection ont été programmés en Pascal sur Vax 785.

Caractéristiques des filtres du VOCODER (SLE) utilisé :		
N° CANAL	BANDE PASSANTE	LARGEUR DE BANDE
1	350	200
2	550	200
3	750	200
4	950	200
5	1 175	250
6	1 450	300
7	1 750	300
8	2 050	300
9	2 350	300
10	2 650	300
11	2 950	300
12	3 300	400
13	3 700	400
14	4 100	400
15	4 700	800
16	5 900	1 600

Figure B2: Bandes passantes du vocoder SLE

CHAPITRE 3

DESCRIPTION SCHEMATIQUE DU PROJET

31. Présentation

Les caractéristiques citées au chapitre 2 nous ont amenés à choisir, pour gérer la variabilité multilocuteur, un système de comparaison dynamique globale de mots isolés avec réduction des données par classification statistique. Comme tout système de reconnaissance, celui-ci possède 2 phases dissociées dans le temps :

- Une phase d'apprentissage (A) faite une seule fois à la mise en place du système pour une application donnée,
- Une phase de reconnaissance (B) correspondant à la partie opérationnelle du système qui est activée à chaque fois qu'un mot est à reconnaitre.

A/ Phase d'apprentissage

Le principe consiste à faire prononcer le vocabulaire de l'application à un grand nombre de personnes, et à choisir parmi toutes ces locutions, celles qui sont les plus caractéristiques, c'est-à-dire les plus représentatives de l'ensemble complet. On utilise pour cela une méthode statistique de réduction de données utilisant seulement les ressemblances de tous ces mots entre eux. A l'issue de cette phase d'apprentissage, le système possède donc un certain nombre de références qui sont stockées sur disque.

B/ Phase de reconnaissance

Lors de cette phase, un mot inconnu est prononcé; pour identifier ce mot le système le compare aux références sélectionnées par l'apprentissage, et l'assimile à celle à qui il ressemble le plus.

La caractéristique essentielle qui distingue ce système (multilocuteur) est donc que toute personne qui se présente au microphone peut être comprise immédiatement sans phase d'apprentissage individuel (comme c'est le cas dans les systèmes monolocuteurs). Que le locuteur fasse ou non partie du corpus d'apprentissage, le système s'adaptera, sauf quelques exceptions, à cette nouvelle voix, et ceci de mieux en mieux au fur et à mesure du dialogue.

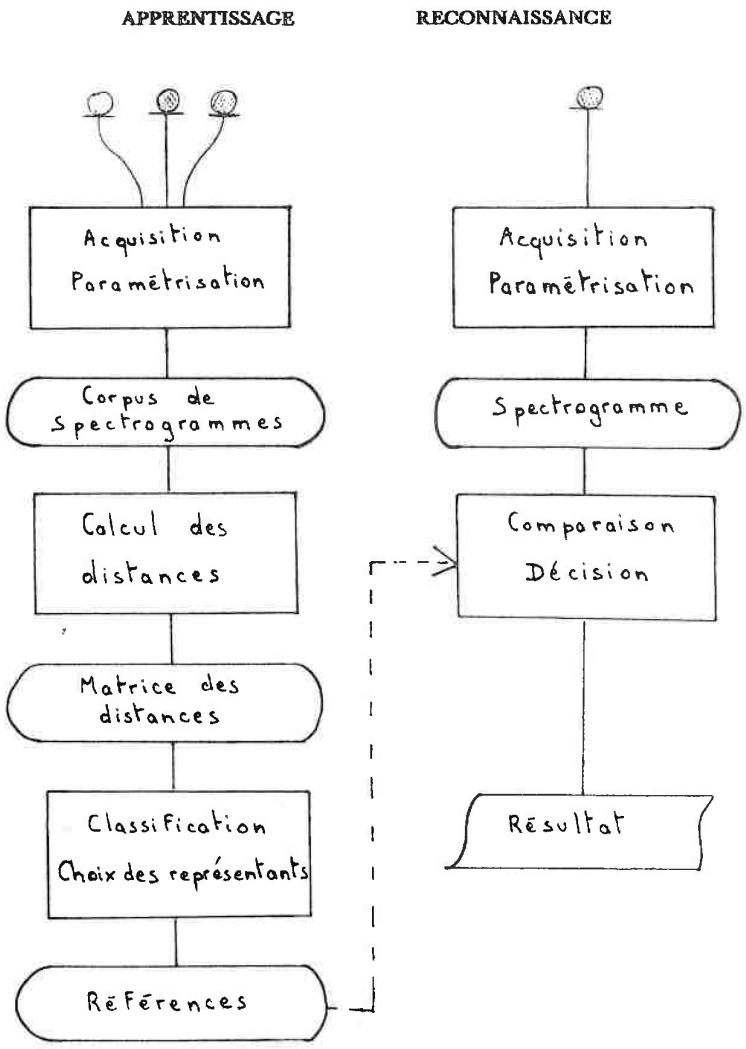
Pour plus de précisions :

- Le passage du signal vocal tel qu'il sort du microphone en un codage exploitable par un programme informatique fait l'objet du paragraphe 43.
- Les paragraphes 44 et 62 détaillent la notion de ressemblance

entre deux mots qui peuvent différer par leur longueur de façon non signifiante (un même mot peut être prononcé vite ou lentement).

* Les méthodes permettant de sélectionner à l'apprentissage les mots les plus représentatifs comme références sont présentées au chapitre A5 pour la théorie et aux paragraphes 45 et 46 pour leurs applications à ce projet.
* La phase de reconnaissance et la façon dont le système s'adapte à un nouveau locuteur inconnu sont développées aux paragraphes 47 et 48.

32. Synoptique



CHAPITRE 4

STRUCTURE DE BASE DU PROJET

41. INTRODUCTION

Ce chapitre détaille la structure de base du projet, c'est-à-dire l'ensemble des étapes et des idées qui sont communes à toutes les variantes testées.

La présentation de ces différentes étapes respecte l'ordre chronologique de leur déroulement réel (cf: synoptique en paragraphe 32) :

- 42 : Apprentissage, reconnaissance
- 43 : Acquisition, paramétrisation
- 44 : Calcul de la matrice des distances
- 45 : Réduction des données par classification
- 46 : Sélection du représentant et du rayon
- 47 : Reconnaissance et décision
- 48 : Adaptation au locuteur.

42. APPRENTISSAGE / RECONNAISSANCE

Ces deux phases, inhérentes à tout système de reconnaissance sont bien distinctes, tant au point de vue de leur utilisation que de leur mise au point. L'apprentissage étant effectué une seule fois lors de la mise en place du système, la priorité y a été donnée au gain de place mémoire plutôt qu'au temps de calcul. Cette phase traite en effet un grand nombre de données complexes sans contraintes de temps.
La phase de reconnaissance, au contraire, est effectuée souvent et traite un nombre restreint de données, mais son utilisation interactive au cours du dialogue nécessite une réponse aussi proche que possible du temps réel ; la priorité est donc pour cette phase le gain de temps, d'où l'importance d'un processeur vectoriel ou de composants spécialisés.

De la taille et de la représentativité du corpus d'apprentissage, dépendent bien entendu les performances ultérieures de reconnaissance. Les algorithmes de classification nécessitent un corpus suffisamment important pour que soient mises en évidence les concentrations de locutions et les prononciations marginales. Bien sûr une prononciation peut être marginale dans un corpus et dominante dans un autre (en raison par exemple des accents régionaux), il est donc important de choisir les locuteurs de la phase d'apprentissage en adéquation avec ceux qui auront à utiliser le système.

Pour les tests nous avons utilisé 20 locuteurs pris au hasard dans les locaux de l'université certains très habitués au microphone d'autres pas du tout. Les statistiques fournies sont établies en séparant ce corpus en deux arbitrairement et de plusieurs façons différentes, moitié pour l'apprentissage, moitié pour la reconnaissance.

On peut distinguer trois types d'applications possibles de ce système en fonction des choix relatifs pour les corpus d'apprentissage et de reconnaissance.

- Applications monolocuteurs:
  Le corpus d'apprentissage est composé de locutions d'une seule personne enregistrées à des moments différents. Cette façon de procéder confère à la reconnaissance sur cette même personne une bonne résistance à la variabilité intra-locuteur et donne des taux très proches de 100%.
- Applications multilocuteurs :
  - Reconnaissance n-locuteurs
    Applicable seulement quand tous les utilisateurs potentiels du système sont connus et en nombre restreint (<100). Les corpus d'apprentissage et de reconnaissance sont alors identiques ce qui rend la classification plus efficace. Les taux de reconnaissance sont intermédiaires entre ceux des applications monolocuteurs et indépendantes du locuteur.
  - Reconnaissance indépendante du locuteur
    C'est plus précisément ce type d'application qui fait l'objet des statistiques et des taux de reconnaissance fournis dans la thèse. Le corpus de reconnaissance est ici différent de celui de l'apprentissage, c'est évidemment le cas le plus délicat. La valeur d'exploitation des références dépend de l'algorithme de classification, de la taille du corpus d'apprentissage et de son adéquation avec l'ensemble des utilisateurs. Cependant, il est certain que ce type de système basé sur des techniques statistiques ne peut manquer d'exclure un faible pourcentage de locuteurs marginaux.

43. ACQUISITION - PARAMETRISATION

Le but de cette étape est de coder sous une forme utilisable par le système les mots qui sont captés par un microphone. Elle est effectuée à l'apprentissage comme à la reconnaissance.

431. Vocodeur

Le signal temporel analogique tel qu'il sort du micro (cf. figure B3) est analysé par un banc de filtres numériques. Il s'agit pour notre application du VOCODEUR SLE décrit au paragraphe 23 qui fournit tous les centièmes de seconde, l'énergie du signal dans seize bandes de fréquences.

Le résultat de cette analyse est un spectrogramme (cf. figure A16) pour lequel :

- Chaque échantillon (c'est à dire chaque ligne) représente un centième de seconde de parole, la dimension temps est donc figurée verticalement.
- Chaque échantillon donne horizontalement les valeurs de l'intensité du signal dans les seize bandes de fréquence de 250 à 6 500 Hz, de gauche à droite ; l'unité est le demi-décibel.

Le premier canal étant peu informant, il a été remplacé, dans notre application, par l'énergie moyenne de l'échantillon :

$$e_1' = \frac{1}{16} \sum_{i=1}^{16} e_i$$

De plus, la visualisation des spectrogrammes est rendue plus lisible grâce au codage de l'intensité en un caractère plus ou moins dense (cf. figures B4 et B7).

- Un programme de détection parole/non-parole basée sur l'énergie du signal (énergie globale et haute fréquence) décide du tri des prélèvement en parole et bruit de fond pour le déclenchement et l'arrêt de l'acquisition de chaque mot prononcé.

Malheureusement, un environnement bruité lors de l'acquisition a rendu cette détection peu fiable surtout en ce qui concerne les fricatives et les plosives non voisées en début et fin de mot. (cf. chapitre 5).

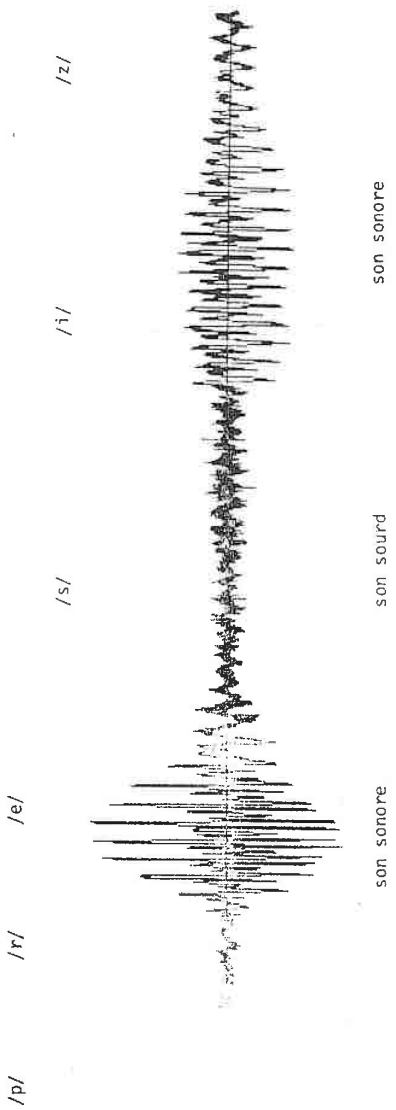


Figure B3: Signal vocal à la sortie du microphone
mot prononcé : ''Précise ''


```
MOT D'APPRENTISSAGE NO:  80
     LOCUTEUR: JEANPAUL
     MOT PRONONCE: CINQ
     SONAGRAMME DE 1190 A 1212
 1                  ::
 2                  ::          /S/
 3                  :+
 4   . . ..      ..+=.
 5  .:::. :. :   .++:
 6  .C3++=+:++::::
 7  +%3C=C=+==++++
 8  =%3C=C=====+=+=
 9  =33C=3===C+=+=.          /ɛ̃/
10  =33C=3C===++++
11  =3C==CC===:++.
12  +CC=====++::..
13  +C=++=++:.:+:
14  +=+:::+++.  ::
15  :=:... .:
16  .:
17
18
19
20
21
22   :C:. .::.              /k/
23  ..+     .    .
```

Figure B4: Spectrogramme codé du mot ''cinq''

432. Codage FFT

Une autre chaîne de paramétrisation plus fiable est étudiée sur un corpus d'apprentissage plus étendu : 4 répétitions de 10 types de locutions proncées par 60 locuteurs (30 hommes et 30 femmes) soit au total un corpus de 2400 locutions: - L'acquisition est réalisée sur Masscomp.
- Le signal est échantillonné à 16 kHz
- L'analyse du signal numérisé est effectuée par un banc de filtres FFT à 32 canaux ayant une répartition linéaire en bandes de fréquence de 0 à 8000 Hz. Le codage étant fait grâce à une fenêtre glissante (Hamming) de 256 points avec un recouvrement de 128 points, on obtient donc un échantillon de parole tous les 0.8 centièmes de seconde.
- On réalise de plus une préaccentuation du signal de 6 dB par octave afin de compenser l'atténuation des énergies hautes fréquences, source principale des problèmes de détection parole/non-parole (cf chapitre 5).

433. Mémorisation

Quelle que soit la méthode utilisée, on obtient, à l'issue de cette phase d'acquisition-paramétrisation lors de l'apprentissage, autant de spectrogrammes numériques que de mots prononcés ; le programme d'acquisition associe à chacun d'eux quelques caractéristiques (cf. figure B5):

- son type : c'est-à-dire le mot qu'il représente,
- le nom du locuteur,
- le nombre de prélèvements.

Toutes ces informations sont stockées sur disque (cf. figure B6) et utilisées lors des phases suivantes.(cf. paragraphes 44 et 45)

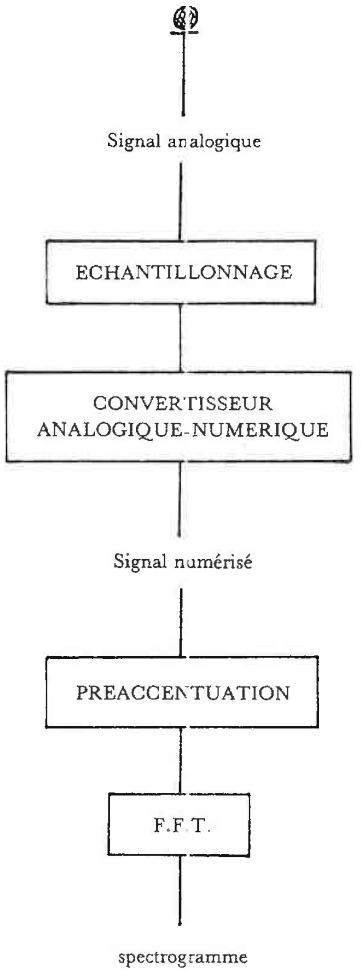


Figure B5: Chaîne d'acquisition avec FFT

```
MOT D'APPRENTISSAGE NO: 237
     LOCUTEUR: MARTINE
     MOT PRONONCE: TROIS
     SONAGRAMME DE 3821 A 3837
  1   :::.:: . .:::
  2  .++::..    ...
  3  .++:.         .
  4  .=+:::.     ..
  5  .=+:..
  6  .++:..
  7  .3C+:+::.....
  8  :%C=+++:::::::.
  9  +%3C=++::++:+.
 10  +3%3C=++:++++:
 11  +3%3C=+++=++=:
 12  +C5%C=++++++++:
 13  +=CCC=++:+:::.
 14  +====+:....
 15  :+:::::.
 16  .:...
 17    . ..
```

```
MOT D'APPRENTISSAGE NO:  81
     LOCUTEUR: JEANPAUL
     MOT PRONONCE: TROIS
     SONAGRAMME DE 1213 A 1228
  1  .=:+:::::.:::..
  2  :+++:. .:  ..
  3  .==+:.:.:..::.
  4  :=:....  .  :.
  5  .3C=++:::..::
  6  :%C=+++:::::++:
  7  +%3===++:+:++:
  8  +%%C==+++++=+::
  9  =%%3==+++====+:
 10  =%%3C=+++==+=+
 11  =%%4C=+++=++++
 12  +3%3C=+++=++++
 13  +CCC=+:::+..::
 14  ++++:.       .
 15  ...:
 16
```

```
MOT D'APPRENTISSAGE NO:  95
     LOCUTEUR: SYLVIE C
     MOT PRONONCE: TROIS
     SONAGRAMME DE 1374 A 1389
  1   ++:::.. ... :.
  2  .+++::       :
  3  .+=+::.      .:
  4  .++:..        .
  5  .==+::       ...
  6  .33C+++:::::+++
  7  +333==+:::+===:
  8  =3C%C=+:::+C==:
  9  =CC%C=+:::+C==:
 10  =CC%C=+:::+C==+
 11  =CC%C=+:::+===+
 12  ===%3=+:::::==+:
 13  +==3C=+::..:..
 14  :++C+:.
 15  .:::+:.
 16
```

```
MOT D'APPRENTISSAGE NO: 165
     LOCUTEUR: ISABELLE
     MOT PRONONCE: TROIS
     SONAGRAMME DE 2598 A 2612
  1   +:::::::. :.::
  2  :+:::::     ...:
  3  .=C=+::. .::::
  4  :=+:::
  5  .+:.
  6   ++:.
  7   C3=++::.:::+:
  8  +C%C=++:::::::+
  9  +C%C=++::+::::+
 10  +=3==+:::+:+:+:
 11  +=3=C+:::+++=+.
 12  +=3C3=+:::++=:
 13  +=3CC=+::::::+:
 14  +=+C+:  .... .
 15  :. ..
```

Figure B6a: Exemple d'une partie du fichier des locutions d'apprentissage

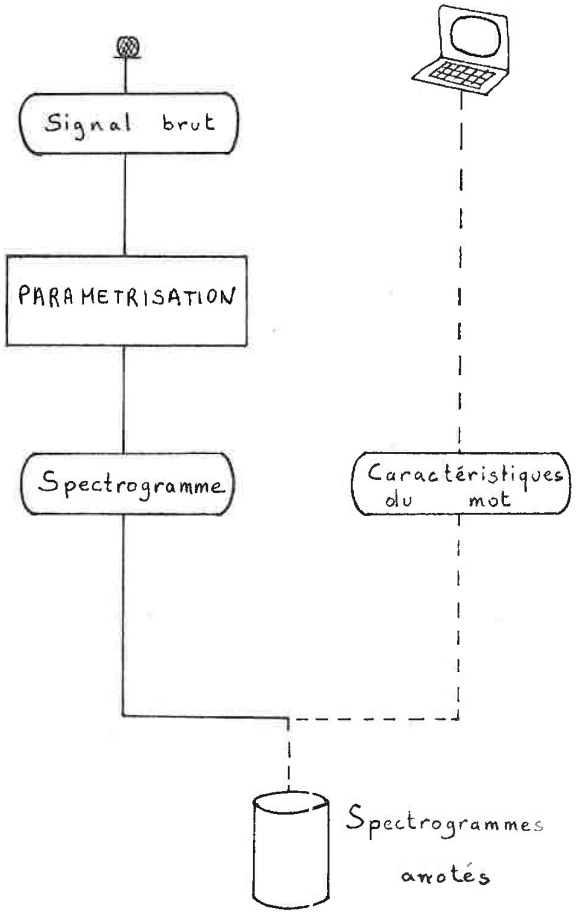


Figure B6b: L'étape d'acquisition-paramétrisation.
(La partie en pointillé n'est effectuée qu'àl'apprentissage)

44. CALCUL DE LA MATRICE DE PSEUDO-DISTANCES

441. But

Le but de cette phase est d'effectuer un pré-traitement sur les spectrogrammes afin d'accélérer l'étape de classification.

Tous les spectrogramme choisis pour la classification (cf. paragraphe 632) sont comparés deux à deux afin de mesurer leur ressemblance. La variabilité de la parole entraîne que deux mots peuvent être structurellement semblables sans être directement comparables, en effet un même mot peut être prononcé rapidement ou lentement, à voix forte ou faible : la figure B7 montre par exemple deux prononciations très différentes du mot "Zéro", l'une faible et rapide l'autre forte et lente. Les difficultés de comparaison soulevées par ces deux types de variations non signifiantes sont traitées dans les 2 paragraphes suivants.

442. Variation de l'énergie des spectrogrammes

Avant d'être comparés entre eux, deux spectrogrammes sont normalisés en énergie. Pour cela on calcule la moyenne des énergies des trois prélèvement les plus forts des deux spectrogrammes. On ramène ensuite à cette valeur moyenne chacun des deux spectrogrammes en multipliant l'énergie de tous leurs prélèvements par le coefficient nécessaire.

La figure B8 montre le résultat de cette normalisation sur les spectrogrammes de la figure B7. Des tests ont été faits sur des corpus avec et sans normalisation de l'énergie, il s'avère que celle-ci a peu d'influence sur les taux de reconnaissance.

443. Variation temporelle

Cette variation est plus délicate car elle n'est pas homogène, en effet non seulement un mot peut être prononcé plus vite par une personne que par une autre mais des distorsions non linéaires de rythme peuvent apparaître, il ne suffit donc pas de ramener les deux formes accoustiques à la même taille pour régler la difficulté : la figure B9a montre deux mots de la même longueur, pourtant dans le premier les phonèmes /s/ et /a/ sont longs et le phonème /y/ est court, dans le second c'est l'inverse, on ne peut donc mettre ces deux mots en correspondance simplement échantillon par échantillon.

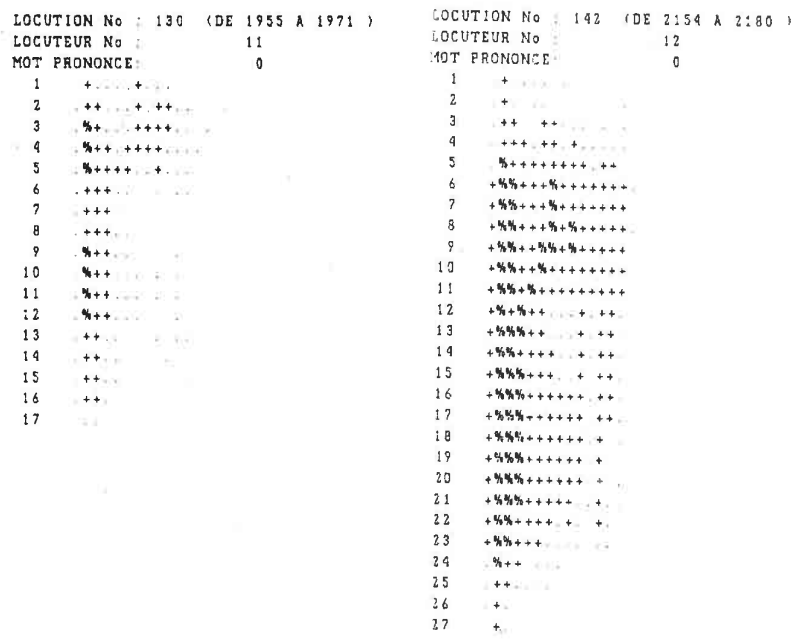


Figure B7: Deux spectrogrammes du mot "zéro"
Le premier est prononcé àvoix faible et rapide,
Le second est prononcé àvoix forte et lente.

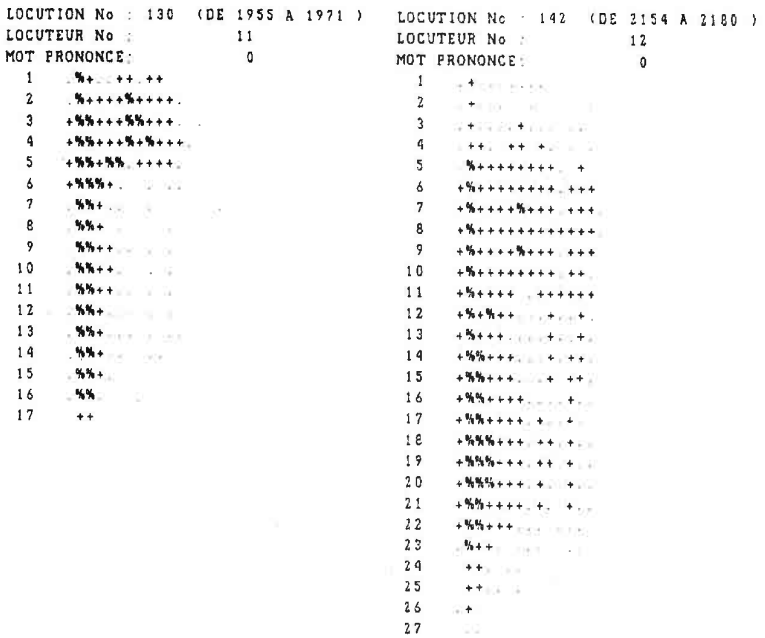


Figure B8: Les deux spectrogrammes de la figure B7 après la normalisation / énergie.

Cette difficulté est levée par l'utilisation pour la comparaison, des algorithmes de programmation dynamique connus depuis longtemps en reconnaissance de mots isolés monolocuteur. Cette technique a pour but la recherche d'un chemin optimal permettant de mettre en correspondance un ou plusieurs prélèvements d'une des deux formes avec un ou plusieurs prélèvements de l'autre (cf figure B 9b).

Nous avons testé pour ce projet deux algorithmes différents dont la description et l'efficacité sont détaillés au paragraphe 62. Les scores issus de ces comparaisons sont d'autant plus faibles que les mots sont plus ressemblants, d'où la notion de distance qu'on peut leur attacher. Nous verrons au paragraphe 624 que toutes les propriétés mathématiques des distances ne sont pas vérifiées, nous appellerons donc "pseudo-distance " le score issu de la comparaison dynamique.

444. Mémorisation

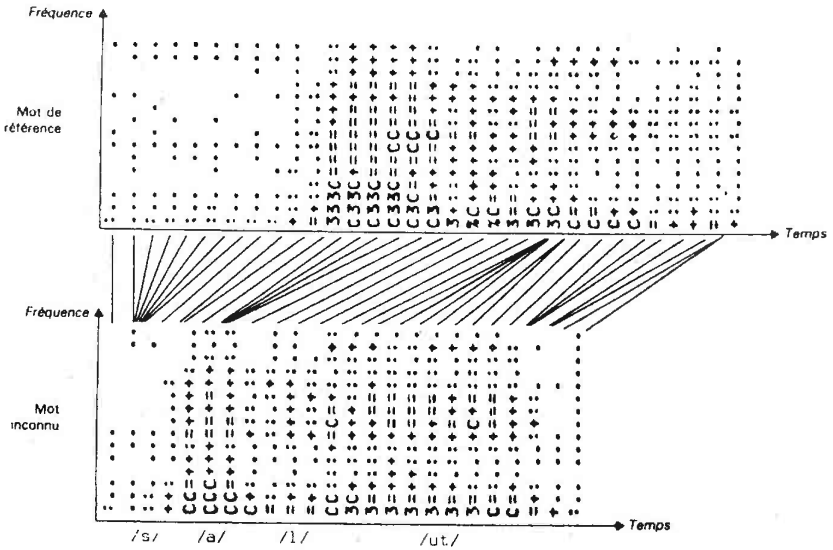
Tous les mots du vocabulaire choisi (cf. paragraphe 632) sont comparés deux à deux, les pseudo-distances obtenues sont rangées dans une matrice triangulaire linéarisée . La figure B10 montre une telle matrice issue de la comparaison de dix occurrences des mots : "un", "cinq" et "trois". Pour plus de clarté,lors de l'impression, les scores supérieurs à un certain seuil ont été remplacés par des étoiles; la matrice réelle contient cependant la totalité des scores.

Si nbloc est le nombre de locutions à comparer, la taille utile de la matrice (c'est-à-dire le nombre de comparaisons) est donnée par :

nbcomp = nbloc * (nbloc - 1)/2

Car on considère acquises les propriétés d'égalité et de symétrie des pseudo-distances. (cf. paragraphe 624).

Le temps de calcul pour 100 locutions (soit 4950 comparaisons) sur microprocesseur MOTOROLA 68000 sans processeur vectoriel est de 45 minutes.



La comparaison de deux mots par programmation dynamique consiste à associer les échantillons successifs des deux mots, sans tenir compte des durées respectives de ces mots. Cette figure montre les correspondances assurées par un tel algorithme sur deux occurrences du même mot « salut » prononcées par un même locuteur sur des rythmes différents. On constate que le recalage temporel effectué permet de compenser les différences de durée des deux mots mais aussi les variations non linéaires de rythme.

Figure B9a:Comparaison dynamique de deux mots [Haton 85]

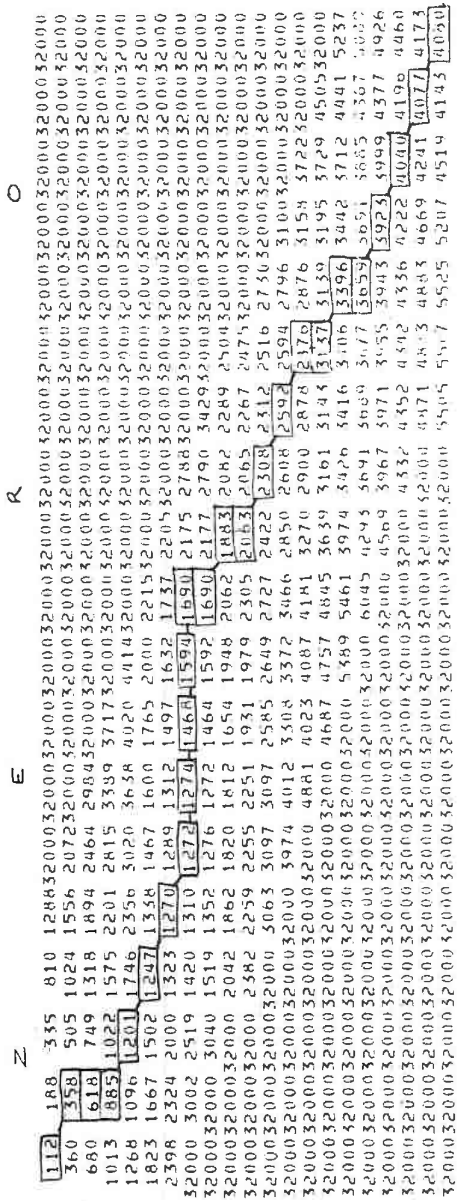


Figure B9b: Comparaison de deux mots avec chemin optimal de correspondance

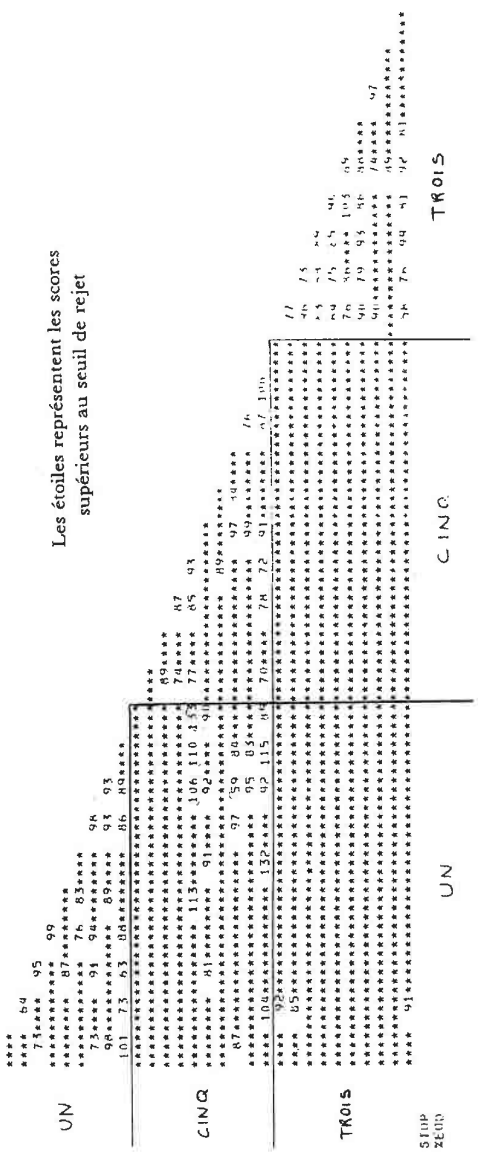


Figure B10: Matrice triangulaire des distances entre les mots ''un'', ''cinq'', ''trois''

45. REDUCTION DES DONNEES PAR CLASSIFICATION

451. But

Conserver la totalité des locutions apprises comme références pour la reconnaissance serait évidemment trop lourd en temps de calcul, surtout pour des corpus importants.

Aussi le but de cette phase et de la suivante (cf. paragraphe 46) est-il de réduire le nombre de données utiles en conservant au mieux l'information contenue dans le corpus d'apprentissage complet. On utilise pour cela un algorithme de classification qui aura pour objectif de répartir un ensemble de N locutions acquises en un nombre p de classes vérifiant les caractéristiques suivantes :

* p << N Le facteur de réduction de N à p dépend de N lui-même, on s'est fixé, pour cette application, une moyenne de 4 ou 5 classes par type de mot.
* Les classes doivent être les plus homogènes possibles (minimisation des distances intra-classes).
* Les classes doivent être aussi différentes les une des autres que possible (maximisation des distances inter-classes).

452. Méthode

Pour ce faire, l'algorithme choisi dispose en entrée :

* De la matrice triangulaire des pseudo-distances entre toutes les locutions (cf. 344).
* D'un tableau contenant pour chaque locution :
  - Le type de mot prononcé
  - Le locuteur
  - Le nombre de prélèvements

Les algorithmes faisant intervenir la position des objets à classer (leurs coordonnées dans un espace) sont donc exclus au profit de ceux n'utilisant que des distances relatives (cf. chapitre A5). Nous avons pour ce projet testé plusieurs méthodes qui diffèrent par :

* Le choix du corpus à classifier. Vaut-il mieux classer l'ensemble du corpus tous types confondus (classification globale), ou au contraire, chaque type de mot séparément (classification par type) ?

Rappelons que le type d'une locution est le mot du vocabulaire qu'elle représente. Exemple : ''UN'', ''NEUF'', ''OUI'' sont trois

des douze types utilisés ici. Les avantages, inconvénients et performances de chaque choix sont discutés au paragraphe 632.

• Le choix de l'algorithme de classification. Nous en avons testé deux : Une méthode proche de UWA, développée par Rabiner aux Bell Laboratories [Rabiner 79], et une méthode de classification ascendante hiérarchique. (cf. paragraphe 633).

• Le choix de diverses variantes:
- Indice d'agrégation, c'est-à-dire définition de la distance d'un point à une classe. Nous avons testé UMI, UMS, DM, Variance. (cf. paragraphe 634)
- Stratégie de segmentation de l'arbre. Une fois les locutions hiérarchisées en un arbre, il reste encore à en couper les branches afin d'obtenir des classes homogènes. Les différentes stratégies de segmentation sont développées au paragraphe 635.
- Une méthode assez différente des autres (CSA) est développée à part au paragraphe 67, en effet, elle ne réduit pas le nombre de données, mais elle les structure afin de gagner du temps à la reconnaissance.

46. SELECTION DU REPRESENTANT ET DU RAYON

461. But

Dans toutes les méthodes présentées au 345 (sauf CSA) le résultat est finalement une partition de l'ensemble de toutes les locutions du corpus d'apprentissage. Chaque sous-ensemble (chaque classe) étant supposé homogène, au moins quant à son type, le but de cette étape est d'en choisir un représentant qui sera seul conservé comme référence de ce type pour la phase de reconnaissance. Là encore plusieurs méthodes sont possibles pour le choix de la locution la plus représentative d'une classe: MINIMAX, MINISOM, POINT MOYEN, elles sont développées au paragraphe 64.

Dans tous les cas on calcule en plus le rayon d'influence de ce représentant, ce qui permet de conserver la notion de taille des classes.

462. Traitement des points isolés

A l'issue de cette étape, nous avons donc autant de formes de référence que de classes obtenues à l'issue de la classification. La ''représentativité'' de chaque référence, c'est-à-dire le nombre de locutions de la classe qu'elle représente, est très variable. Certaines sont des singletons correspondant à des points isolés qui n'ont que peu d'intérêt. C'est pourquoi avant leur mémorisation définitive, nous faisons subir aux références un traitement qui élimine les points isolés les moins significatifs. (cf. paragraphe 65) Toutes les étapes décrites des paragraphes 43 à 46 constituent la phase d'apprentissage, et ne sont donc effectuées qu'une fois à la mise en place du système. Le résultat de cette phase est une liste de références (en moyenne quatre ou cinq par type de mot) qui sont conservées sur disque avec pour chacune d'elle :

- Son spectre
- Son type
- Son rayon d'influence.

Ces informations sont les seules qui sont utilisées pour la reconnaissance.(cf figure B11).

```
MOT DE REFERENCE NO: 11
     NUMERO DU MOT(ZZPOIN) : 15
     LOCUTEUR: PASCAL D
     MOT PRONONCE: SIX
     SONAGRAMME DE  218 A  234
     RAYON:167
     TYPE:  5

MOT DE REFERENCE NO: 12
     NUMERO DU MOT(ZZPOIN) : 75
     LOCUTEUR: JEANPAUL
     MOT PRONONCE: SIX
     SONAGRAMME DE 1107 A 1131
     RAYON:115
     TYPE:  5

MOT DE REFERENCE NO: 13
     NUMERO DU MOT(ZZPOIN) :123
     LOCUTEUR: MOHAMED
     MOT PRONONCE: SIX
     SONAGRAMME DE 1835 A 1859
     RAYON:115
     TYPE:  5

MOT DE REFERENCE NO: 14
     NUMERO DU MOT(ZZPOIN) :183
     LOCUTEUR: LAURENT
     MOT PRONONCE: SIX
     SONAGRAMME DE 2900 A 2912
     RAYON:152
     TYPE:  5

MOT DE REFERENCE NO: 15
     NUMERO DU MOT(ZZPOIN) :207
     LOCUTEUR: FLORENCE
     MOT PRONONCE: SIX
     SONAGRAMME DE 3312 A 3317
     RAYON:170
     TYPE:  5

MOT DE REFERENCE NO: 16
     NUMERO DU MOT(ZZPOIN) :124
     LOCUTEUR: MOHAMED
     MOT PRONONCE: QUATRE
     SONAGRAMME DE 1860 A 1878
     RAYON:127
     TYPE:  6

MOT DE REFERENCE NO: 17
     NUMERO DU MOT(ZZPOIN) : 88
     LOCUTEUR: SYLVIE C
     MOT PRONONCE: QUATRE
     SONAGRAMME DE 1306 A 1330
     RAYON:130
     TYPE:  6
```

```
MOT D'APPRENTISSAGE NO: 123
     LOCUTEUR: MOHAMED
     MOT PRONONCE: SIX
     SONAGRAMME DE 1835 A 1859
 1                  :::=
 2   .     .    . .:+=
 3  ..           .:+=
 4  ..         . ..:==
 5  ...    ....  ::==
 6  ..     :     .::==
 7  .:     . .   .:+:
 8  .=+++:+===:....:
 9  :=++:+===+:::.
10  +=+++:++=++:::.
11  :=++::+=++:::
12  :=+++::+=+:.:.
13  :=+:::+=++.::.
14  :::.. ....  .::
15  ..           :.:+
16                .  :
17                .  +
18  .              :+
19  ..       -     .+=
20  .      .  .    :+=
21               .  .:+=
22  .              .:=
23                 ..++
24                 ..::
25                 .
```

Figure B11: Une partie du fichier des références

47. RECONNAISSANCE ET DECISION

471. But

Cette étape représente la phase opérationnelle du système, elle est effectuée chaque fois qu'un mot inconnu est prononcé et elle a pour but de classer cette locution parmi les types connus ou de la déclarer inclassable.

472. Méthode

Dans un premier temps, le mot prononcé est paramétrisé de la même manière que lors de l'apprentissage (cf. paragraphe 43), on obtient alors un spectrogramme qui est ensuite comparé séquentiellement à toutes les références à l'aide du même algorithme que celui utilisé lors de la création de la matrice de pseudo-distances (cf. paragraphe 44). On décide alors :

- soit que le mot prononcé est du type de la référence qui lui ressemble le plus,
- soit, si les scores de comparaison sont trop grands, que le mot est inconnu.

Remarquons qu'on ne décrète pas d'ambiguité (en cas de scores trop proches) puisque de toutes façons on demande confirmation du mot reconnu.

Nous avons testé plusieurs critères permettant d'attribuer une locution à une référence en fonction des scores obtenus :

* critère du plus proche voisin (P.P.V.) (ne tient pas compte du rayon d'influence)

* critère du plus proche voisin à l'intérieur du rayon.

* critère de ''l'intériorité''.

Ces différents critères ainsi que leur efficacité sont détaillés au paragraphe 66.

La classification conservant la structure d'arbre (CSA) requiert évidemment une méthode de reconnaissance différente. La recherche de la meilleure référence se fait non plus séquentiellement mais par exploration dichotomique d'un arbre (cf. paragraphe 67).

Dans tous les cas on conserve les deux meilleures réponses afin de pouvoir proposer une deuxième solution en cas de non-confirmation de la première (voir l'organisation du dialogue figure B1).

48. ADAPTATION AU LOCUTEUR

481. But

Puisque nous avons choisi de faire confirmer par le locuteur chaque mot prononcé, nous pouvons mettre à profit cet avantage pour adapter le système au locuteur courant au fur et à mesure de la poursuite du dialogue.

482. Méthode

Le principe consiste à construire progressivement un lexique de références propres au locuteur parallèlement au lexique général construit par l'apprentissage (cf. figure B1). Pour cela, à chaque fois qu'un mot a été confirmé par le locuteur, son spectrogramme est stocké dans le lexique particulier de ce locuteur ; ainsi, incomplet au départ, il s'enrichit progressivement à chaque confirmation.(cf. figure B12).

Quand un mot inconnu est prononcé il est comparé sans distinction aux deux lexiques fusionnés.

Le lexique particulier ne conservant (au plus) qu'une référence par type de mot, c'est à chaque fois, la dernière locution prononcée qui est mémorisée.

Cette adaptation au locuteur est légèrement coûteuse en temps de reconnaissance, car elle ajoute une référence supplémentaire à chaque type de mot. Pour les normes que nous nous sommes fixées (quatre ou cinq références par type) le temps de reconnaissance est multiplié en moyenne par 1.2. Elle est cependant intéressante dans deux types d'application :

- Si le nombre de prononciations de chaque mot est grand, donc dans le cas d'une utilisation prolongée du système,
- Si le lexique particulier peut être mémorisé et repris lors d'une utilisation ultérieure par le même locuteur, ce qui implique une identification du locuteur, un vocabulaire assez restreint et un nombre d'utilisateurs potentiels pas trop important.

Cette adaptation a été implantée dans le cadre du premier projet (MULOT) et aboutissait à une amélioration sensible du taux de reconnaissance après quelques répétitions.

Des statistiques précises sont malheureusement impossibles car nous ne disposons pas encore d'un nombre suffisant de répétitions du vocabulaire par un même locuteur.

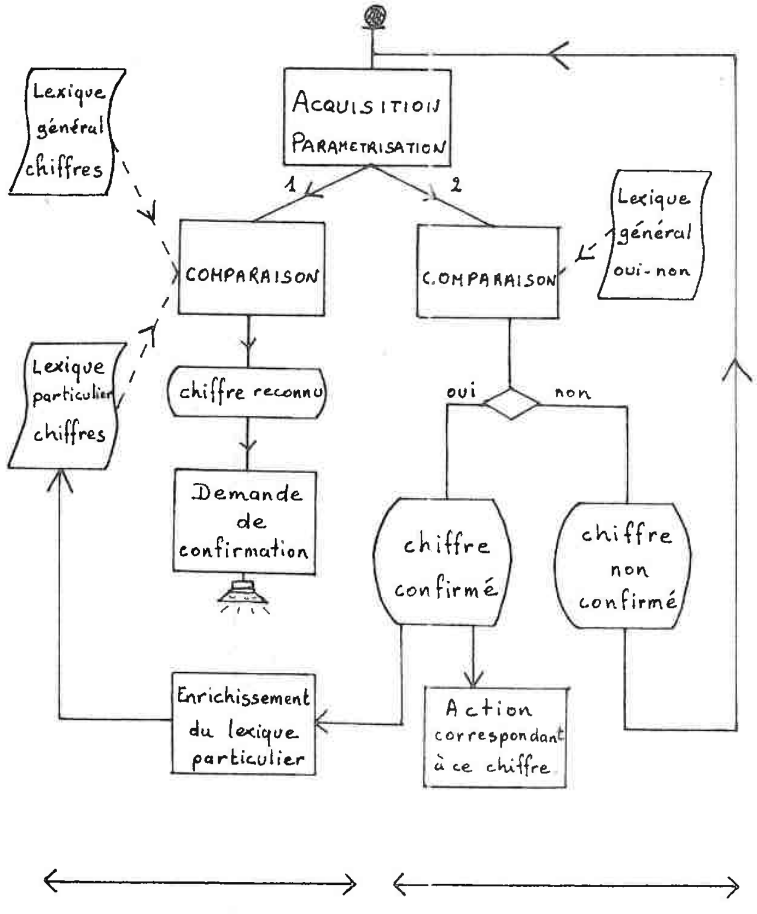


Figure B12: Organisation du système d'adaptation au locuteur

CHAPITRE 5

CARACTERISTIQUES DU CORPUS

L'ensemble des locutions prononcées au cours de la phase d'apprentissage et mémorisées sous forme de spectrogrammes (corpus d'apprentissage) possède certaines caractéristiques qu'il est utile de détailler ici car elles influent grandement sur les choix et les statistiques présentés au chapitre suivant.

51. Problèmes de détection parole non-parole

Nous avons vu au paragraphe 43 qu'un environnement bruité a rendu la détection parole/non-parole peu fiable. Ceci est particulièrement visible en ce qui concerne les fricatives et les plosives non voisées en début et fin de mot qui sont trop peu énergétiques pour déclencher ou maintenir la procédure d'acquisition. La figure B13 par exemple montre un ''UN'' et un ''CINQ'' correctement acquis et un ''CINQ'' où manquent la sifflante /s/ du début et la plosive /k/ de la fin, la confusion avec ''UN'' est pratiquement inévitable. La fréquence de cette anomalie fait des chiffres français un vocabulaire difficile, en effet ''CINQ'', ''SIX'', ''SEPT'', ''HUIT'' présentent cette configuration ainsi que ''QUATRE'' quand il est prononcé /kat/ ce qui est assez fréquent. Une telle diminution des caractéristiques de chaque mot du vocabulaire rend la classification moins homogène (par exemple pour ''six'' il faut au moins quatre classes ''SIS'', ''SI'', ''IS'' et ''I'') et les confusions lors de la reconnaissance plus fréquentes (par exemple ''CINQ'' et ''UN'' sur la figure B13).

Si l'on classifie à part les deux sous-vocabulaires :
''ZERO'', ''UN'', ''DEUX'', ''TROIS'', ''QUATRE'',
et ''CINQ'', ''SIX'', ''SEPT'', ''HUIT'',''NEUF'' ,
on constate que le second, qui regroupe presque toutes les locutions difficiles, fournit des résultats de reconnaissance très sensiblement moins bons que le premier (77% / 97% contre 93% / 100%).

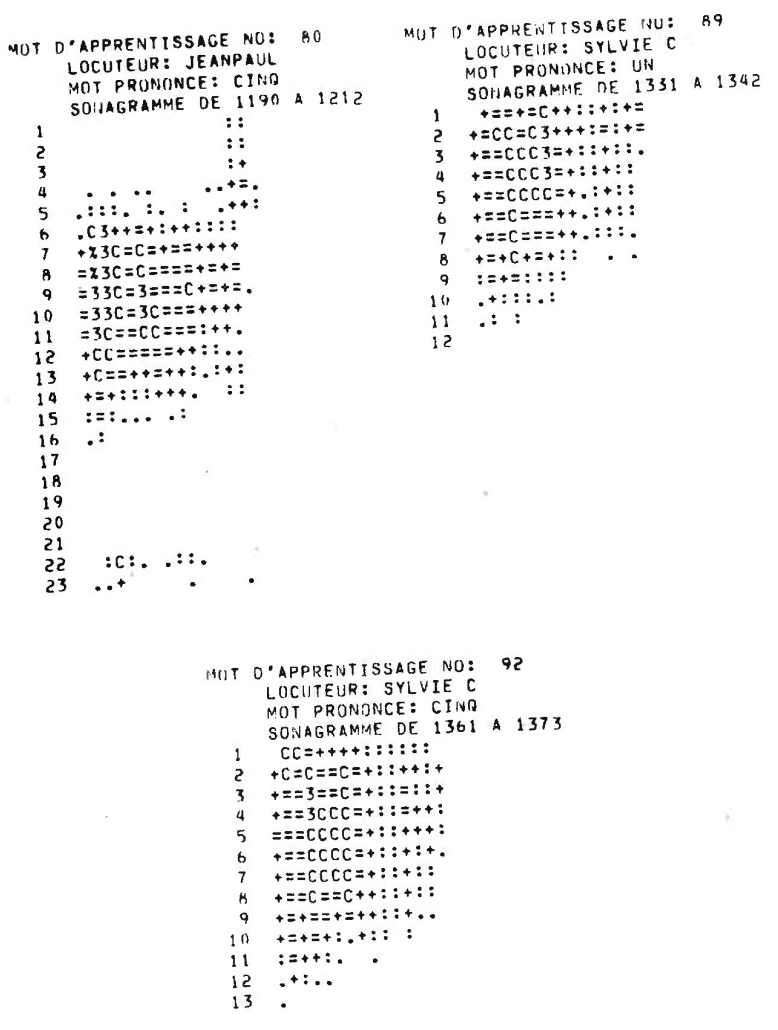


Figure B13: Trois spectrogrammes illustrant les problèmes de détection parole / non-parole

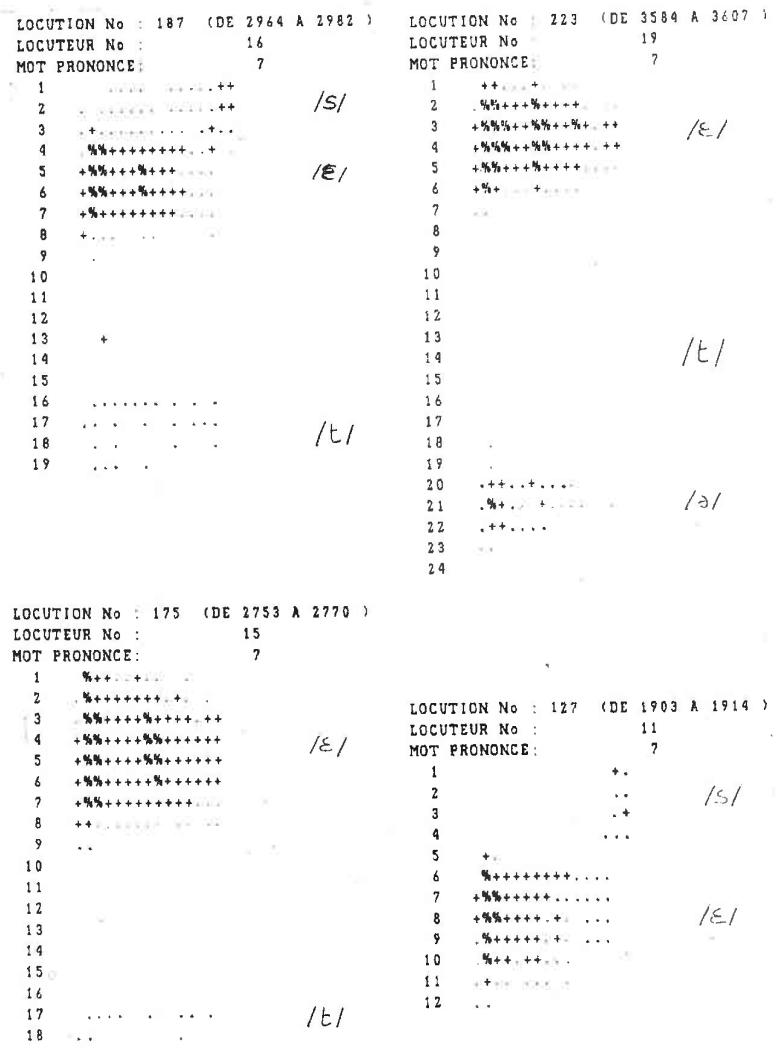


Figure B14: Quatre spectrogrammes différents correspondant au même mot: ''sept''

52. Répartition des locutions suivant le type

Rappelons que le type d'un mot est sa signification. On constate sur le corpus d'apprentissage que la répartition des locutions à l'intérieur d'un même type est très variable, on peut classer les critères de répartition en trois dimensions :

- densité
- homogénéité
- différenciation

• Types denses / lâches

La densité d'un type peut se mesurer par inversement au diamètre (c'est à dire la plus grande distance intra-type) et au rayon (c'est à dire la distance du centroide au point le plus éloigné). Ce critère est important lors de la phase de représentation, en effet la notion de centroide n'a pas la même valeur selon la densité de la classe qu'il représente. On voit sur la figure B16a que les densités peuvent varier pratiquement du simple au double selon les types, et on verra au paragraphe 64 qu'il s'est avéré intéressant de conserver cette information lors de la reconnaissance sous forme d'un ''rayon d'influence '' associé à chaque centroide. On constate également sur cette figure que les types les plus lâches sont les types polymorphes (voir ci-dessous).

• Types homogènes / polymorphes

Certains types peuvent présenter plusieurs formes très différentes, c'est le cas de ceux possédant une fricative ou une plosive en début ou fin de mot pour les raisons exposées au paragraphe 51. Par exemple, ''SEPT'' peut présenter les formes suivantes /set/, /se/, /setə/ , /et/ (cf. figure B14). Evidemment les types polymorphes nécessitent pour leur représentation plus de références que les types homogènes.

Remarque : l'homogénéité est une forme de densité. Les types polymorphes ont en effet une distance intra-type importante, mais chacune des formes de ce type peut être dense ou lâche. (cf. figure B15).

• Types différenciés / empiétants

Ce critère mesure la distance inter-types (cf figure B15). Il est évidemment fondamental pour les étapes de classification (cf. figure B17) et de reconnaissance . On voit sur la figure B16b que ''TROIS'' et ''SIX''

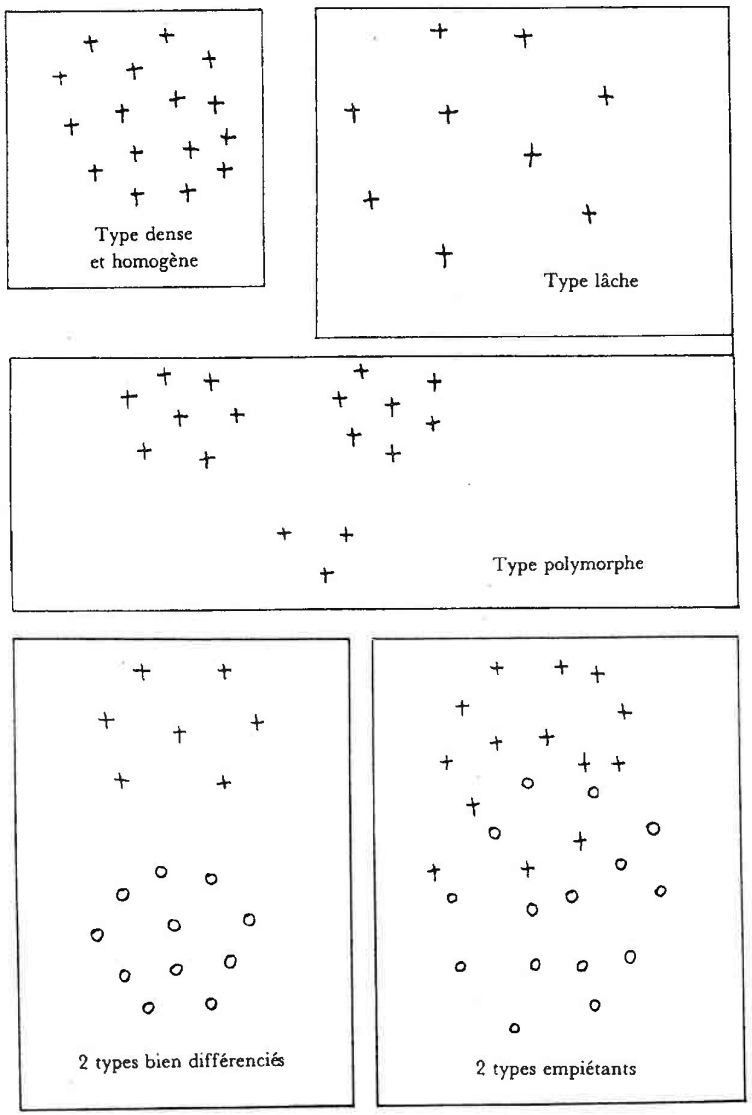


Figure B15: Représentation schématique des différents critères de répartition des types

par exemple, sont des types bien différenciés des autres au contraire de "QUATRE " et "SEPT". Lors de la reconnaissance, on peut s'attendre à des confusions entre "QUATRE ", "SEPT", "CINQ", "HUIT" (locutions terminées par une plosive) et entre "DEUX" et "ZERO".
Lors de la classification, les types bien différenciés ("TROIS" par exemple) seront facilement séparés, donc il suffira de peu de références pour les représenter, inversement les types "SEPT" et "CINQ" très empiétants, entraîneront des difficultés lors de la segmentation, avec pour conséquence, une multiplication des références (cf. figure B17) et des singletons ainsi que des confusions fréquentes lors de la reconnaissance.

Toutes ces caractéristiques influeront sur le choix des méthodes et des paramètres présentés au chapitre 6.

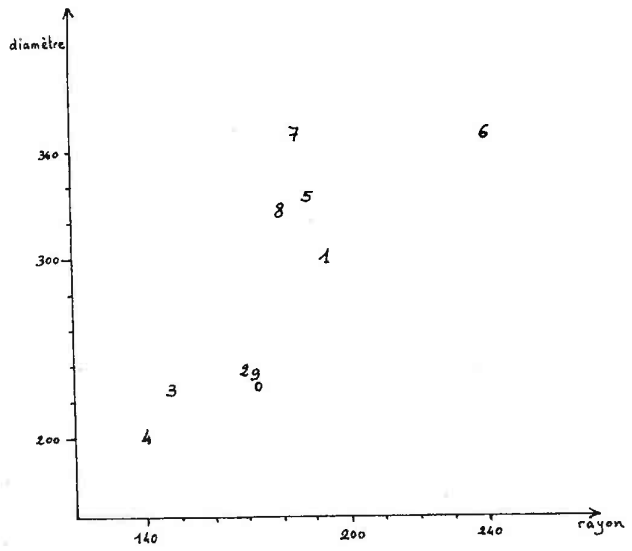


Figure B16a: Répartition des types en fonction de leur densité

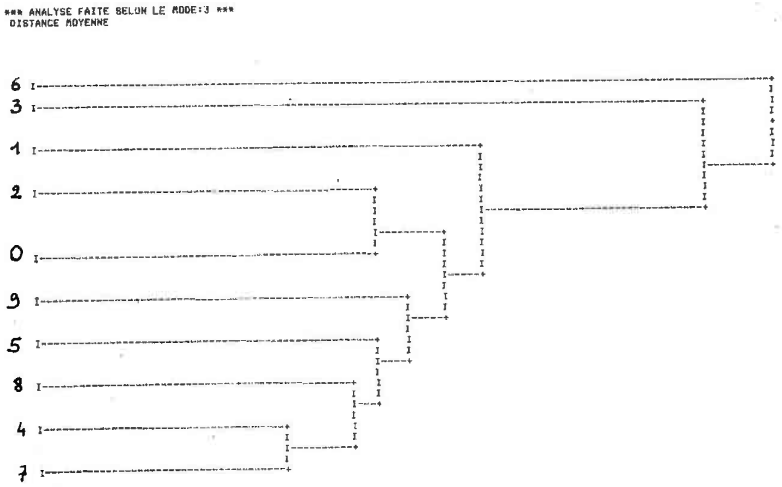
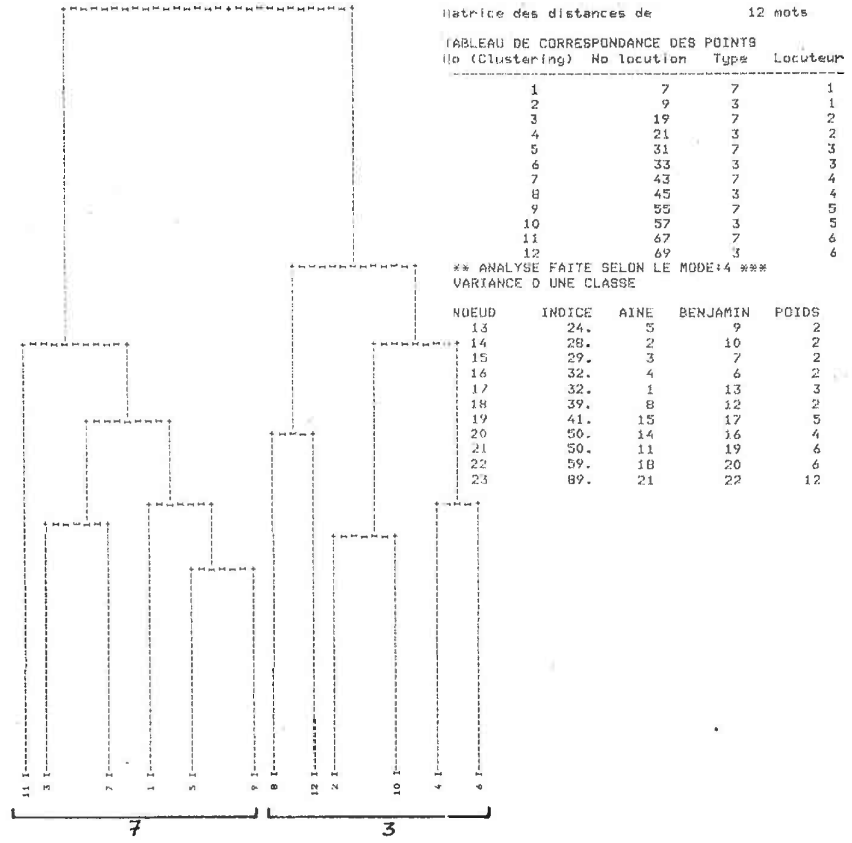


Figure B16b: Répartition hiérarchique des types selon leurs distances respectives



RESULTATS DU CLUSTERING

Coefficient d'homogeneite pour un cluster : 1.00

CLUSTER DE TYPE 7:

Numerotation clustering:	11	3	7	1	5	9
Numerotation ds TLOCUT :	67	19	43	7	31	55
Apres elim. des P.I. :	67	19	43	7	31	55

CENTRE DU CLUSTER: 1(7)
RAYON: 146.00

CLUSTER DE TYPE 3:

Numerotation clustering:	8	12	2	10	4	6
Numerotation ds TLOCUT :	45	69	9	57	21	33
Apres elim. des P.I. :	45	69	9	57	21	33

CENTRE DU CLUSTER: 4(21)
RAYON: 143.00

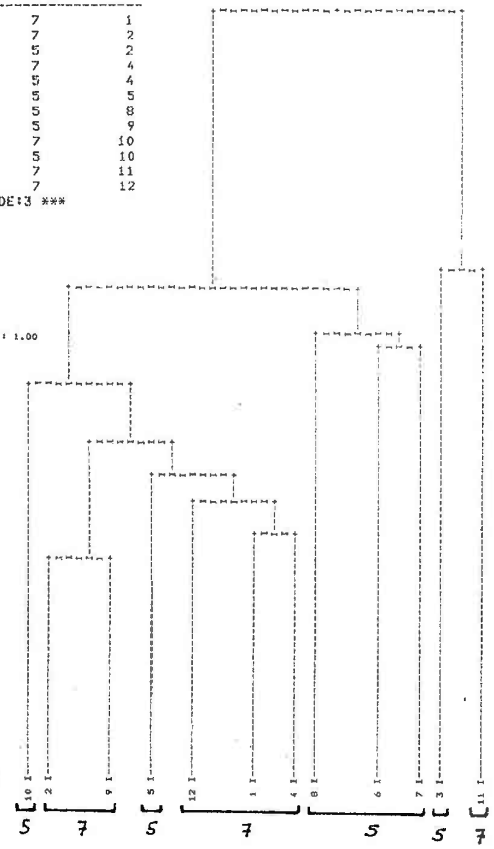
Figure B17a: Arbre et classification de deux types bien différenciés

Matrice des distances de 12 mots

TABLEAU DE CORRESPONDANCE DES POINTS

No (Clustering)	No locution	Type	Locuteur
1	7	7	1
2	19	7	2
3	20	5	2
4	43	7	4
5	44	5	4
6	56	5	5
7	92	5	8
8	104	5	9
9	115	7	10
10	116	5	10
11	127	7	11
12	139	7	12

*** ANALYSE FAITE SELON LE MODE:3 ***
DISTANCE MOYENNE



RESULTATS DU CLUSTERING

Coefficient d'homogeneite pour un cluster : 1.00

CLUSTER DE TYPE 5:
Numerotation clustering: 10
Numerotation ds TLOCUT : 116
Apres elim. des P.I. : 116

CENTRE DU CLUSTER: 10(116)
RAYON: 0.00

CLUSTER DE TYPE 7:
Numerotation clustering: 2 9
Numerotation ds TLOCUT : 19 115
Apres elim. des P.I. : 19 115

CENTRE DU CLUSTER: 2(19)
RAYON: 70.00

CLUSTER DE TYPE 5:
Numerotation clustering: 5
Numerotation ds TLOCUT : 44
Apres elim. des P.I. : 44

CENTRE DU CLUSTER: 5(44)
RAYON: 0.00

CLUSTER DE TYPE 7:
Numerotation clustering: 12 1 4
Numerotation ds TLOCUT : 139 7 43
Apres elim. des P.I. : 139 7 43

CENTRE DU CLUSTER: 4(43)
RAYON: 80.00

CLUSTER DE TYPE 5:
Numerotation clustering: 8 6 7
Numerotation ds TLOCUT : 104 56 92
Apres elim. des P.I. : 104 56 92

CENTRE DU CLUSTER: 7(92)
RAYON: 140.00

CLUSTER DE TYPE 5:
Numerotation clustering: 3
Numerotation ds TLOCUT : 20
Apres elim. des P.I. : 20

CENTRE DU CLUSTER: 3(20)
RAYON: 0.00

CLUSTER DE TYPE 7:
Numerotation clustering: 11
Numerotation ds TLOCUT : 127
Apres elim. des P.I. : 127

CENTRE DU CLUSTER: 11(127)
RAYON: 0.00

Figure B17b: Arbre et classification de deux types empiètants

CHAPITRE 6

DESCRIPTION DETAILLEE DES CHOIX ET DE LEUR INFLUENCE

61. INTRODUCTION

Autour de la structure de base du projet détaillée au chapitre 4 nous avons testé plusieurs variantes et mesuré leur efficacité quant au taux de reconnaissance.

611. Paramètres dépendants

Précisons tout de suite que pour toutes les comparaisons les taux de reconnaissance sont donnés toutes choses optimales par ailleurs (et non toutes choses égales par ailleurs) en effet nous ne pouvons pas exclure à priori la dépendance de deux ou plusieurs paramètres. Par exemple, sur la figure B18, si nous testons toutes choses égales par ailleurs, d'une part le choix B/B' (1), d'autre part le choix D/D' (2) nous concluons que B est meilleur que B' et D meilleur que D', pourtant la chaîne testée avec B' et D' ensemble (3) peut donner un meilleur résultat. (Les paramètres B et D sont alors dépendants).

C'est pourquoi nous avons testé toutes les combinaisons possibles et donnons les résultats à l'optimum. Sur cet exemple nous donnerions comme résultat pour le choix B/B':

choix B toutes choses optimales par ailleurs (c'est-à-dire avec choix D) --->85%
choix B' toutes choses optimales par ailleurs (c'est-à-dire avec choix D') ---> 87%.

A chaque fois que ce cas de figure se présentera (ce qui s'est révélé assez rare) nous le préciserons.

612. Nombre de références

Les taux de reconnaissance sont liés au nombre de références conservées par l'apprentissage (qui peut varier de 40 à 50 pour l'ensemble des 10 chiffres). Les taux qui sont fournis dans ce chapitre sont comparés à nombre de références autant que possible identique.

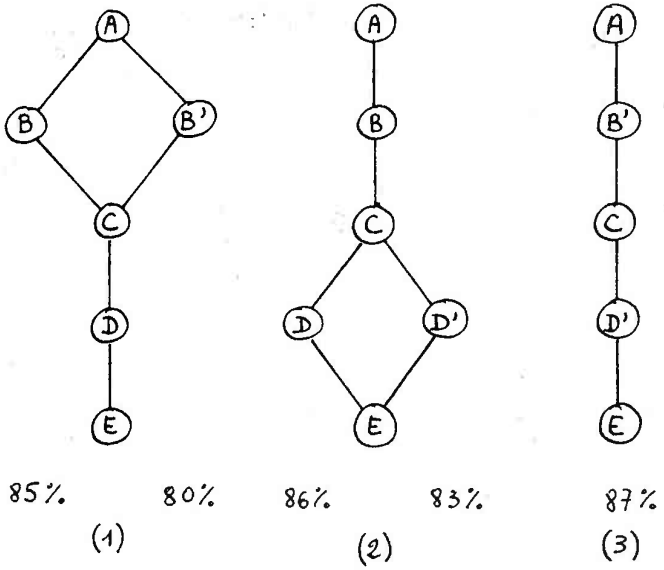


Figure B18: Illustration de la dépendance des paramètres

613. Corpus de test

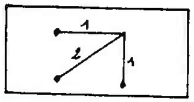
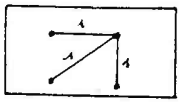
Ces statistiques concernent le vocabulaire des chiffres en français. Les taux fournis sont des moyennes, l'apprentissage étant fait avec dix locuteurs, les tests de reconnaissance avec dix autres, et ceci de plusieurs manières.

614. Rappel

Rappelons également que les résultats des statistiques comportent deux pourcentages :

* l'un (le plus faible nécessairement) concerne le taux de reconnaissance exacte lorsque le système ne propose qu'une réponse ;
* l'autre, lorsque le système propose deux réponses possibles. (cf. algorithme de dialogue).

615. Synoptique des différentes variantes

Etape	choix	Variantes				paragr.
ACQUISITION PARAMETRISATION						§ 43
CALCUL DE LA MATRICE DES DISTANCES	Algorithme de programmation dynamique					§ 62
REDUCTION DES DONNEES	Choix du corpus	Global		par type		§ 632
	Méthode	non-hiérarchique		C.A.H.		§ 633
	Indice d'agrégation	UMI	UMS	D.M.	Var.	§ 634
	Stratégie de Segmentation	à niveau fixe	à niveau variable guidée par le nb de points max par classe	à niveau variable guidée par le type des feuilles	Pas de segmentat. (C.S.A.)	§ 635 et § 67
CHOIX DU CENTROÏDE ET DU RAYON	CENTROÏDE	Minimax	Mini som	Point moyen		§ 64
TRAITEMENT DES SINGLETONS	POINTS ISOLES	laissés	traités	supprimés		§ 65
RECONNAISSANCE	STRATEGIE DE RECONNAISSANCE	P.P.V.	Rayon	Intériorité		§ 66

62. LES ALGORITHMES DE PROGRAMMATION DYNAMIQUE

621. But

Rappelons que les algorithmes de programmation dynamique permettent de comparer deux formes telles que décrites au paragraphes 43, en calculant une pseudo-distance ; cette opération intervient à deux étapes du système :

- Lors du calcul de la matrice des pseudo-distances où tous les mots du corpus d'apprentissage sont comparés deux à deux (cf. paragraphe 44),
- Lors de la reconnaissance où le mot prononcé inconnu est comparé aux références issues de l'apprentissage (cf. paragraphe 47).

Il est bien évident que quel que soit l'algorithme choisi c'est le même qui est utilisé pour les deux étapes.

622. Principe de la programmation dynamique

a) Présentation

La programmation dynamique est une méthode de résolution des problèmes d'optimisation basée sur le principe d'optimalité locale de Bellman [Bellman 57]. Son application aux problèmes de reconnaissance de forme a déjà été largement traitée, principalement dans le domaine monolocuteur [Sakoe 71].

Son principe consiste à rechercher l'association optimale des échantillons successifs des deux formes, en contractant ou dilatant pas à pas l'axe temporel de sorte que la distance entre les deux formes soit minimale (cf. figure B19 et B20).

b) Formalisation

Considérons les deux formes à comparer A et B ayant été paramétrisées par un vocodeur à n canaux, elles sont chacune représentées par une suite d'échantillons, c'est-à-dire de vecteurs de $\mathbb{R}^n$, la longueur de cette suite peut être très variable puisqu'elle est fonction du temps mis à prononcer le mot.

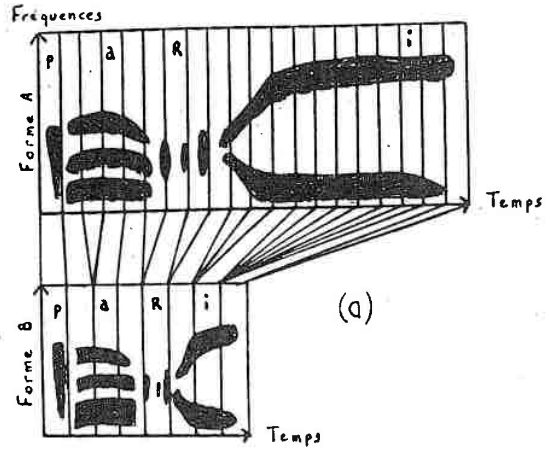


Figure B19: Distorsion temporelle sur deux occurrences du mot ''Paris''

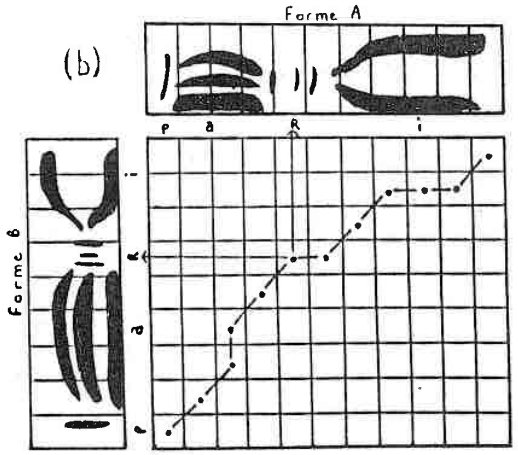


Figure B20: Chemin de recalage optimal obtenu par programmation dynamique [Haton 79]

Soit I le nombre d'échantillons de A
et J le nombre d'échantillons de B, on a
$A = \{a1, a2, \ldots, ai, \ldots, aI\}$ $ai \in \mathbb{R}^n$
$B = \{b1, b2, \ldots, bj, \ldots, bJ\}$ $bj \in \mathbb{R}^n$

La mise en correspondance de deux vecteurs est fondée sur une notion de distance minimum. Parmi les nombreuses définitions possibles, (Hamming, Itakura ...) nous avons choisi la distance de Minkovski déterminée par :

$$d(a, a') = \sum_{k=1}^{n} | P_k - P'_k |$$

où $a = (P1, P2, \ldots Pn)$ et $a' = (P'1, P'2, \ldots, P'n)$

Soit le plan (X,Y), considérons deux axes X et Y représentant la suite temporelle des échantillons de A et B respectivement.

La distance D(A,B) sera déterminée en fonction d'un chemin dans le plan (X,Y) allant du point (1,1) au point (I,J).

Ce chemin peut être décrit par la suite de ses étapes :

$C = (c(1), c(2), c(3), \ldots, c(K))$

avec $\mathrm{Max}(I,J) \leq K \leq I + J$

c(k) étant le couple (i(k), j(k))

où i et j sont les indices des échantillons respectifs de A et B mis en correspondance à l'étape k (cf. figure B21).

Le chemin ainsi défini doit vérifier certaines contraintes :

* Conditions aux limites
  $i(1) = j(1) = 1$
  $i(K) = I$
  $j(K) = J$

Le chemin commence par la mise en correspondance des deux premiers échantillons et finit par la mise en correspondance des deux derniers de

chaque forme.

- Condition de continuité
  $i(k) - i(k-1) \leq 1$
  $j(k) - j(k-1) \leq 1$

Tous les échantillons de chaque forme sont mis en correspondance au moins une fois (pas de saut).

- Condition de monotonie
  $i(k-1) \leq i(k)$
  $j(k-1) \leq j(k)$
  $(i(k-1), j(k-1)) \neq (i(k), j(k))$

Le chemin de recalage suit le cours du temps dans les deux formes (pas de retour arrière).

On détermine pour cela une équation de programme dynamique qui définit récursivement une distance locale sur ce chemin :

$$g(c(k)) = g(i,j)$$

et une distance globale normalisée entre les deux formes

$$D(A,B) = g(I,J) / L(I,J)$$

où L(I,J) est la longueur du chemin de recalage.

Deux fonctions g ont été testées pour ce projet :

- Première équation de programmation dynamique (C- COMPARE)

$$g(i,j) = \text{Min} \begin{cases} g(i,j-1) + d(i,j) \\ g(i-1,j-1) + 2d(i,j) \\ g(i-1,j) + d(i,j) \end{cases}$$

avec $L(I,J) = I+J$ (cf. Figure B22)

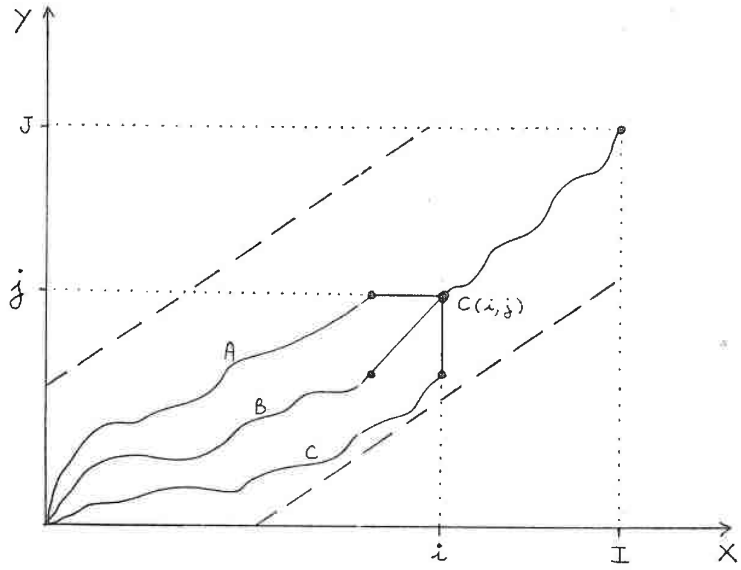


Figure B21: Le chemin de recalage en (i,j) est l'optimum des trois chemins A, B et C, eux-mêmes optimum de l'étape précédente

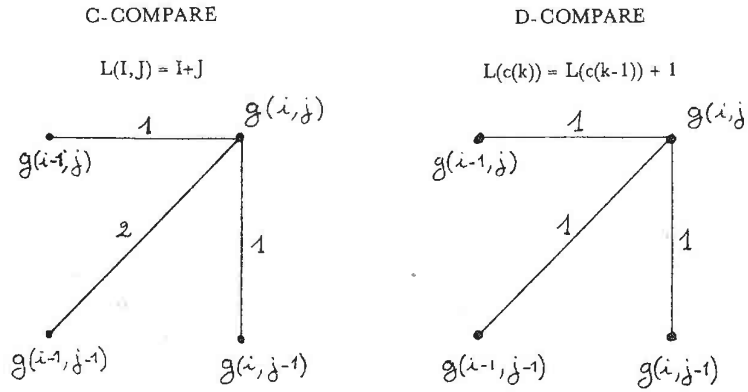


Figure B22: Les deux équations de programmation dynamique testées

• Deuxième équation de Programmation dynamique (D-COMPARE)

$$g(i,j) = \min \left\{ \begin{array}{l} g(i,j-1) \\ g(i-1,j-1) \\ g(i-1,j) \end{array} \right\} + d(i,j)$$

En choisissant comme longueur le nombre d'étapes du chemin

$$L(I,J)$$

définit récursivement par

$$L(c(k)) = L(c(k-1)) + 1 \quad \text{(cf. Figure B22)}$$

Deux algorithmes (C-COMPARE et D-COMPARE) correspondantes à ces deux équations ont été testés et ne donnent pas de différences sensibles quant aux taux de reconnaissance.

Pour plus de précision sur les différentes méthodes, nous renvoyons le lecteur à [Sakoe 78] ou [Haton 82] et pour une comparaison plus générale sur l'influence des contraintes sur les taux de reconnaissance à [di Martino 81].

623. Rejets

• Lors de la phase de reconnaissance seulement, un système d'interruption prématurée de la comparaison est effectué dans deux cas :

- Rejet sur longueur : si les deux mots à comparer sont trop différents en nombre d'échantillons, la comparaison dynamique n'est pas faite, le score rendu est un infini conventionnel. Le rapport maximal des deux longueurs nécessaires pour commencer la comparaison est un paramètre du système. Nous avons fixé cette valeur lors des tests à 2.5.

- Rejet sur score : l'utilisateur du système peut donner un score maximum, au delà duquel on n'accepte pas que deux mots soient assimilés. C'est également un paramètre du système qui mesure sa sévérité. Si ce seuil est trop bas de nombreux mots corrects seront rejetés, s'il est trop haut des mots n'appartenant pas au vocabulaire seront reconnus à tort. Il existe donc une procédure d'interruption qui intervient au cours de la comparaison dès qu'on

est sûr que le score final sera supérieur au seuil autorisé. Le score rendu est alors un autre infini conventionnel.

624. Notion de pseudo-distance

Les scores issus de cette comparaison, sont d'autant plus faibles que les mots sont plus ressemblants, d'où la notion de distance qu'on peut leur attacher. Nous avons testé les caractéristiques de ces scores par rapport aux propriétés classiques des distances mathématiques.

1- Positivité : $d(x,y) \geq 0 \quad \forall x, \quad \forall y$

2- Egalité : $d(x,x) = 0 \quad \forall x$

3- Symétrie : $d(x,y) = d(y,x) \quad \forall x, \quad \forall y$

4- Inégalité triangulaire : $d(x,y) \leq d(x,z) + d(z,y) \quad \forall x,y,z.$

Ici, x, y, z sont les spectrogrammes représentant les mots du vocabulaire et d(x,y) représente le score issu de la comparaison dynamique des spectrogrammes x et y. Pour les deux algorithmes retenus, les trois premières propriétés sont vérifées dans tous les cas. La quatrième propriété présente quelques exceptions :

Pour cent spectrogrammes comparés deux à deux il y a (100*99*98)/2 triangles soit 485100 inégalités à vérifier.

• Pour le premier algorithme (C-COMPARE) on observe un dépassement dans :
91 cas soit 0,02%
dont 34 cas dépassent de plus de 5% du score soit 0,01%.

• Pour le second algorithme (D-COMPARE) on observe un dépassement dans :
257 cas soit 0,05%
dont 127 cas de plus de 5% du score soit 0,02%.

Ces mêmes tests sur d'autres matrices donnent des taux très semblables. Contrairement aux résultats obtenus dans des conditions

approchantes par E. Vidal-Ruiz [Vidal-Ruiz 85], nous observons donc ici une légère violation de l'inégalité triangulaire toutefois, la très faible proportion d'exceptions nous autorise à parler de pseudo-distances pour les scores issus de la comparaison dynamique.

Tout au long de ce projet, nous utiliserons implicitement des notions topologiques telles que classes de proximité, centroïde, rayon, PPV, etc... Ces notions peuvent être considérées comme pertinentes dans la mesure où l'espace et la métrique envisagés respectent pratiquement les propriétés des distances.

63. REDUCTION DES DONNEES

631. Introduction

L'étape de réduction des données par classification donne lieu à plusieurs choix importants que nous allons envisager dans l'ordre chronologique de leur apparition :

- choix du corpus à classifier,
- choix de la méthode de classification,
- choix d'une notion de distance pour l'indice d'agrégation des classes,
- Stratégie de segmentation de l'arbre hiérarchique.

632. Choix du corpus à classifier

Rappelons que le type des objets à classifier (c'est-à-dire la signification de chaque spectrogramme) est connu à priori, d'où deux idées possibles pour tenir compte de cette information :

a) Classification par type

Il s'agit ici de classifier les mots de chaque type séparément. Si N est le nombre de locuteurs du corpus et p le nombre de types de mots dans le vocabulaire, on fait p classifications de N formes. Les références obtenues pour chaque type sont ensuite rassemblées en un seul fichier.

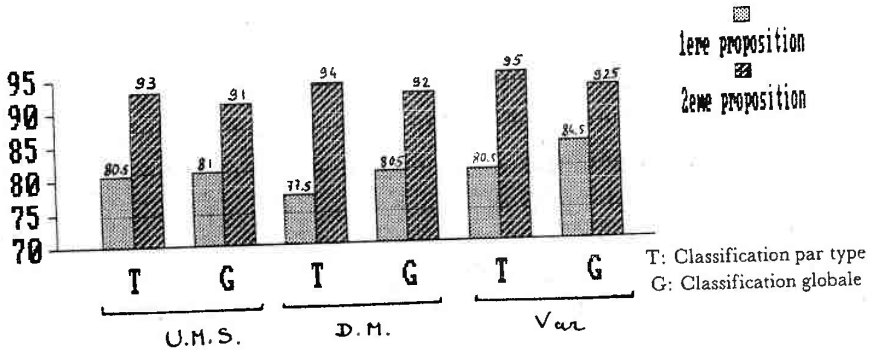
b) Classification globale

Tous les mots du corpus d'apprentissage sont classifiés en une seule fois tous types confondus. On effectue une classification de N x p formes "guidée " par l'information sur les types. Nous verrons au paragraphe 635 de quelle manière cette information peut guider la segmentation de l'arbre en classes.

On constate sur la figure B23 que la classification globale donne des résultats qui sont meilleurs en première proposition et légèrement moins bons en deux propositions. L'optimum est :

Pour la classification globale (avec 43 références) : 84.3 - 92.5.
Pour la classification par type (44.5 références) 80.5 - 95.

Dans les deux cas, il a lieu lorsqu'on choisit la variance (cf. paragraphe 634) comme critère d'agrégation. De plus la classification par type s'est révélée plus irrégulière au cours des tests. La classification



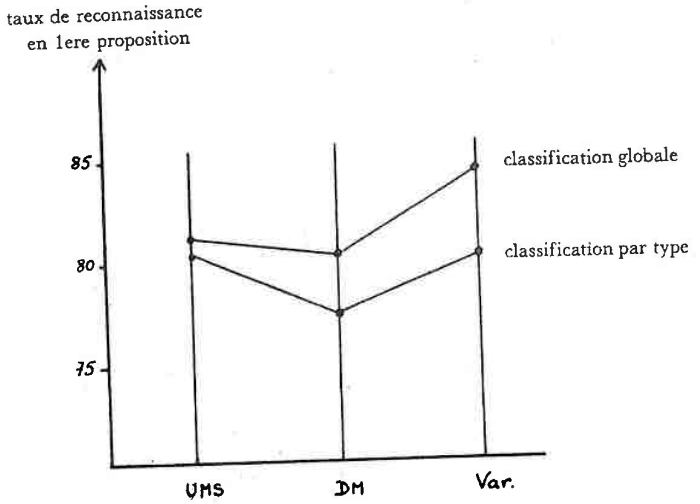


Figure B23: Comparaison des taux de reconnaissance pour les classifications globale et par type

globale a donné des résultats plus constants donc plus fiables, les optima des différents paramètres testés étant contenus dans une fourchette plus restreinte.

Cette légère supériorité de la classification globale sur la classification par type peut s'expliquer par une meilleure information sur l'interpénétration des classes (des tests faits sur le corpus d'apprentissage montrent que le taux de recouvrement inter-type est de 6.5% pour la classification globale contre 11.5% pour la classification par type). En effet dans la classification par type chaque découpage en classes est fait sans tenir compte de la répartition des formes des autres types (cf. figure B24). Le critère de séparation sera dans ce cas un paramètre du système (nombre de points maximum par classes, nombre de classes) identique pour tous les types (cf. paragraphe 635) ce qui rend mal compte des différences d' homogénéité des types : nous avons vu au chapitre 5 que certains types (''TROIS'' par exemple) sont très bien différenciés et peuvent donc être représentés par peu de références (1 ou 2), d'autres au contraire sont empiétants ou/et polymorphes (''SEPT'', ''CINQ'', ...) et nécessitent plus de références (jusqu'à 8 quelquefois) pour rendre compte des diverses prononciations. (cf. figure B15 chapitre 5).

Nous verrons au paragraphe 635 et 64 comment la classification globale peut prendre en compte ces différences d'homogénéité des classes. L'inconvénient de la classification globale est sa lourdeur, dans le cas de notre corpus l'ensemble à classer est 10 fois plus grand pour la méthode globale que pour la méthode par type ; les calculs intervenant pour la matrice des distances est donc plus de 100 fois plus importants, ce qui rend la méthode globale inapplicable aux très grands corpus.

633. Choix de la méthode

Parmi les nombreuses méthodes statistiques de classification (cf. Partie A chapitre 5) qui n'utilisent en entrée qu'un tableau de distances, nous avons vu qu'on pouvait distinguer deux catégories: les méthodes hiérarchiques et non hiérarchiques.

Nous avons testé une méthode non hiérarchique et une méthode de classification ascendante hiérarchique (CAH).

a) Méthode non hiérarchique

Cette méthode, est issue de la méthode UWA [Rabiner 76] présentée en partie A, paragraphe 522, et enrichie principalement pour tenir compte de

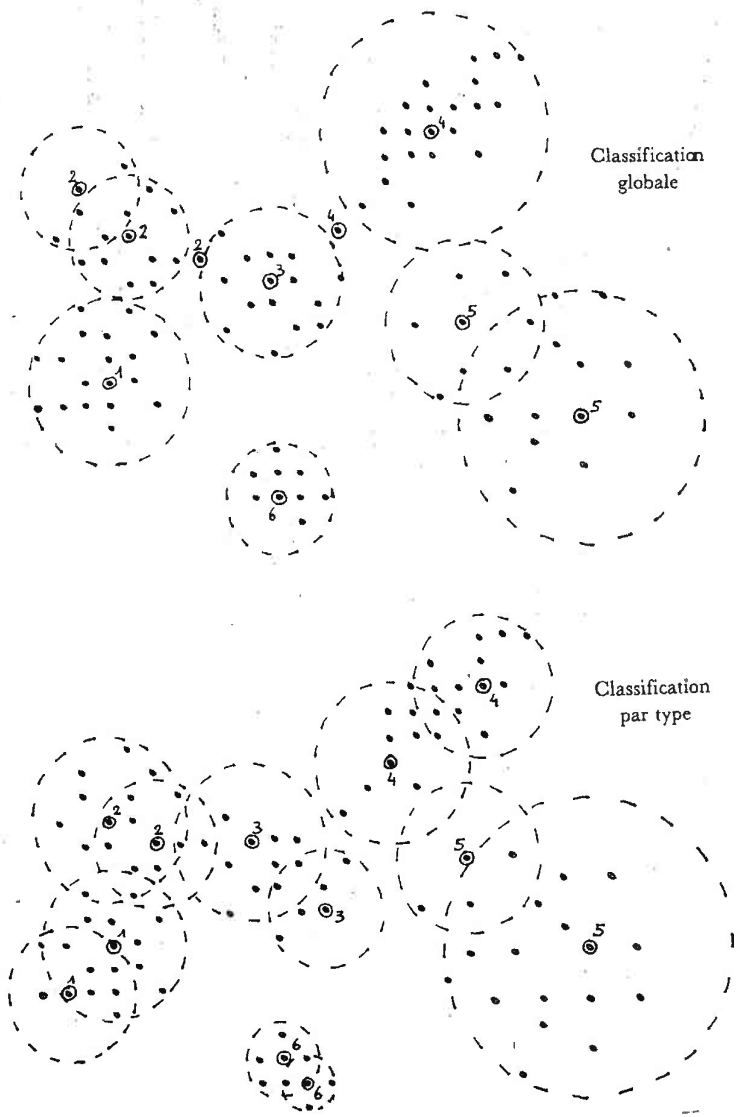


Figure B24: Influence du choix : Global / Par type
sur l'empiètement des classes (10 références)
(Les chiffres représentent les types associés aux centroides).

l'information sur les types des feuilles lors de la classification globale. La description détaillée de cette méthode se trouve dans [Divoux 82]. Nous en donnons ici un résumé :
L'algorithme part de l'ensemble complet des points-locutions à classer : E Il peut s'agir, nous l'avons vu au paragraphe précédent, soit de l'ensemble de toutes les locutions du corpus d'apprentissage (méthode globale), soit seulement des locutions de même type (méthode par type). Au fur et à mesure que l'algorithme définit des classes, il en calcule un représentant et un rayon et les retire de l'ensemble E jusqu'à obtention de l'ensemble vide.

Détermination d'une classe :
Rappel : chaque point représente une locution. A partir du centroide de l'ensemble E le programme délimite une sphère w de points (son rayon est un paramètre du système) dont on calcule le centroide, généralement différent mais voisin du précédent. L'opération est réitérée jusqu'à obtention d'un ensemble w stable (c'est-à-dire identique sur deux étapes consécutives). Si la définition du centroide est bien choisie (cf. notion de centroide MINISOM paragraphe 64), le programme détermine ainsi un chemin de proche en proche jusqu'à une région stable de densité maximale (cf. figure B25).

Lorsque l'ensemble w est stable, un centroide représentant en est conservé comme référence ainsi qu'un rayon d'influence (cf. paragraphe 64). Les points situés à l'intérieur de cette sphère d'influence forment une classe et sont retirés de l'ensemble E . Le processus complet est alors réitéré jusqu'à obtention de l'ensemble E vide.

Remarque: Lors de la classification globale, l'information sur les types intervient à deux niveaux:

- Lors de la détermination des classes, l'algorithme de recherche des zones de forte densité ne tient compte que du type majoritaire dans le voisinage.
- Lors du calcul du "rayon d'influence " (cf paragraphe 646): Une fois le centroide (de type T) déterminé, le rayon est incrémenté régulièrement jusqu'à la dernière locution de type T (RMIN), puis jusqu'à la première locution de type différent (RMAX); le rayon d'influence est donné par la moyenne de RMIN et RMAX.

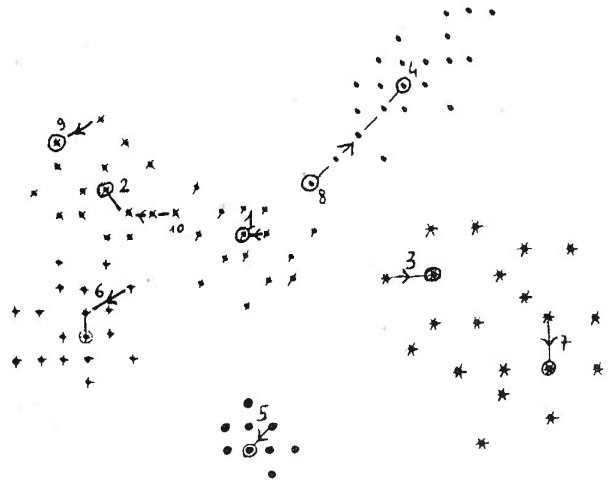


Figure B25a: Classification sur un nuage de 100 points de 6 types différents.
On constate le chemin parcouru par les centroides avant de se stabiliser dans les zones de forte densité. (Les chiffres représentent l'ordre dans lequel ils ont été trouvés).

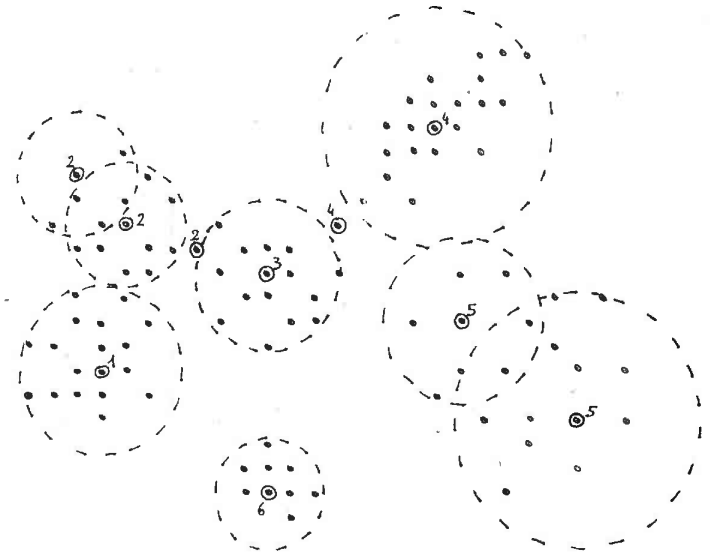
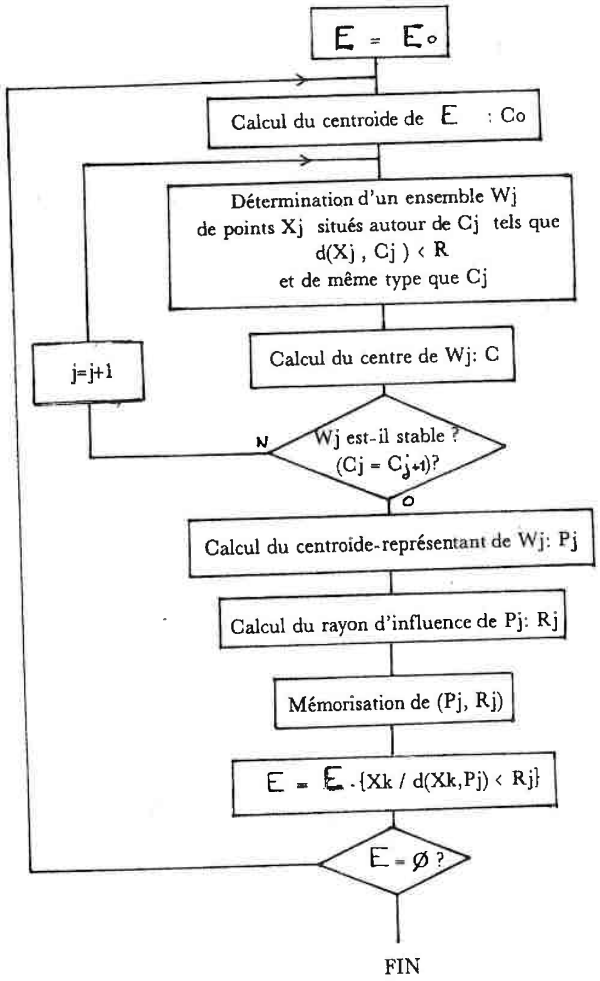


Figure B25b: Centroides définitifs et sphères d'influences.
On constate qu'il n'y a pas de recouvrement de sphères de types différents.
(Les chiffres représentent les types associés aux centroides).

Organigramme de la classification non-hiérarchique

(Voir l'algorithme en partie A, paragraphe 522)



b) Classification ascendante hiérarchique (CAH)

Nous avons vu (Partie A, chapitre 5) que lors d'une classification ascendante hiérarchique la partition de l'ensemble de toutes les formes de départ en un nombre plus restreint de classes, se fait en deux phases :

- construction ascendante de l'arbre représentant la hiérarchie des formes.
- segmentation de l'arbre en branches représentant les classes.

Première phase: Construction de l'arbre:

La classification ascendante hiérarchique procède par agrégations successives des classes les plus proches à chaque étape.

* A l'étape 0 il y a N classes comportant chacune une seule locution, N étant la taille du corpus d'apprentissage (10 ou 100 pour notre application, suivant qu'il s'agit d'une classification par type ou globale (cf. paragraphe 632)). La matrice D des pseudo-distances de toutes ces formes entre elles a été pré-calculée (cf. paragraphe 44).

On a donc :

$H_o = \{C1, C2, \ldots Cn\}$

avec $\forall j, Cj = \{lj\}$

$D = D_o$

où Cj sont les classes et lj les locutions.

* A l'étape i pour i de 1 à N-1

On cherche parmi les N-i+1 classes de l'étape précédente : C1, C2,...Cn-i+1 ; les deux classes les plus proches Ca et Cb qu'on agrège en une seule nouvelle classe CaUb = Ca U Cb. Hi = {C1, C2,.... Cn-i}

* On calcule la nouvelle matrice des pseudo-distances Di en recalculant les N-i-1 distances des classes autres que Ca et Cb à la nouvelle classe Caub, les autres valeurs de la matrice restant inchangées.

* A l'issue de la dernière étape il ne reste plus qu'une classe réunissant l'ensemble des locutions.

$H_{n-1} = \{C1\}$ $C1 = \{l1, l2, \ldots ln\}$

Nous avons vu au paragraphe 62 comment on définissait la distance entre deux locutions ; l'algorithme d'agrégation fait ici intervenir la notion de proximité entre classes de locutions ; nous verrons au paragraphe 634 les différentes méthodes pour passer de l'une à l'autre.

L'ensemble du processus d'agrégation constitue une hiérarche de partitions qui peut être visualisée par un arbre (cf. figure B26) dans lequel :

- Les locutions du corpus de départ sont représentées en bas par un numéro,
- Les noeuds créés lors de la classification sont numérotés de N+1 à

Matrice des distances de 10 mots

TABLEAU DE CORRESPONDANCE DES POINTS

No (Clustering)	No locution	Type	Locuteur
1	3	6	1
2	15	6	2
3	27	6	3
4	39	6	4
5	51	6	5
6	63	6	6
7	75	6	7
8	87	6	8
9	99	6	9
10	111	6	10

*** ANALYSE FAITE SELON LE MODE:4 ***
VARIANCE D UNE CLASSE

NOEUD	INDICE	AINE	BENJAMIN	POIDS
11	31.	2	4	2
12	36.	5	9	2
13	37.	6	8	2
14	43.	1	13	3
15	48.	3	10	2
16	61.	7	11	3
17	66.	12	14	5
18	82.	15	16	5
19	126.	17	18	10

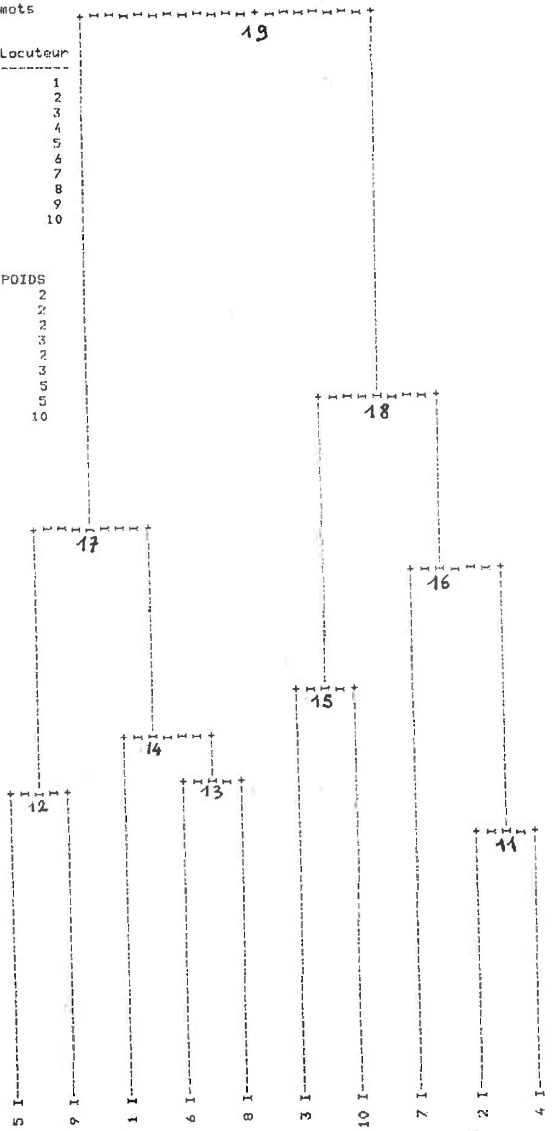


Figure B26: Arbre et tableau représentant la CAH de 10 énonciations du mot ''six''

2N-1 où N est le nombre de locutions à classer.
- Chaque étape est représentée par un segment horizontal reliant deux locutions ou deux classes ou une locution et une classe,
- La hauteur de ce segment est proportionnelle à l'indice d'agrégation c'est-à-dire à la distance des deux classes. Plus deux locutions sont proches, plus tôt elle sont reliées.

Deuxième phase : Segmentation de l'arbre

La première phase se contente de structurer les locutions en fonction de leur ressemblance ; la deuxième a pur but d'isoler des classes de locutions. Pour cela, un algorithme va ''débiter'' l'arbre en un certain nombre de branches dont les feuilles constitueront autant de classes. Les différentes stratégies de segmentation sont envisagées au paragraphe 635, elles ont en commun de donner un découpage en classes toutes disjointes.

c) Résultats

On constate sur la figure B27 que la classification hiérarchique parait mieux adaptée que la classification UWA; en effet, avec des résultats en 2 propositions à peu près équivalents (1.5 % de différence) la CAH donne de bien meilleurs résultats en une seule proposition (+ 8%) et avec 43 références aulieu de 51. Il est à noter que ces deux méthodes sont comparables quant à la quantité d'informations conservées relatives au corpus d'apprentissage. (Des tests de reconnaissance effectués sur le corpus ayant servi à l'apprentissage donnent: Pour UWA, 90%-95% et pour CAH 91%-96%). Il semble donc que la CAH conserve une information plus pertinente, c'est à dire plus générale, moins spécifique du corpus d'apprentissage.
Elle présente cependant l'inconvénient d'être légèrement plus lourde en temps de calcul que UWA.

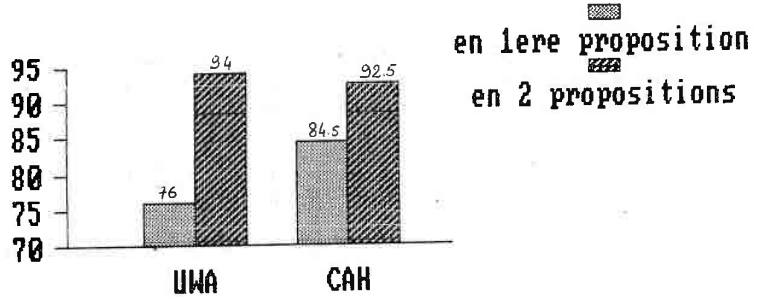


Figure B27: Comparaison des taux de reconnaissance selon le type de classification: nodale (UWA) ou hiérarchique (CAH)

634. Choix d'une distance pour l'indice d'agrégation

Ce choix ne concerne que la classification ascendante hiérarchique ; il a pour but de trouver une méthode permettant de calculer, à partir de la distance entre locutions définie au paragraphe 62, une distance entre classes de locutions. Quatre mesures différentes dont les définitions ont été données au paragraphe 532, ont été testées pour ce projet :

- Ultramétrique inférieure (UMI)
- Ultramétrique supérieure (UMS)
- Distance moyenne (DM)
- Variance (VAR)

Rappelons qu'à chaque étape, deux classes seulement sont agrégees et que les formules de récurrence démontrées dans [Jambu 78] (cf paragraphe 532) permettent de calculer la matrice de l'étape i uniquement en fonction de celle de l'étape i-1.

L'exemple schématique suivant (figures B28 a, b et c) montre sur un corpus de cinq points l'évolution de la matrice et l'arbre hiérarchique équivalent. Les figures B29 a et b montrent la même évolution sur un échantillon réel de neuf locutions réparties en trois types.
N.B. : A chaque étape la distance minimum entourée détermine les deux classes à agréger.

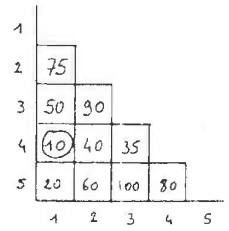


Figure B28a: Matrice de départ précalculée

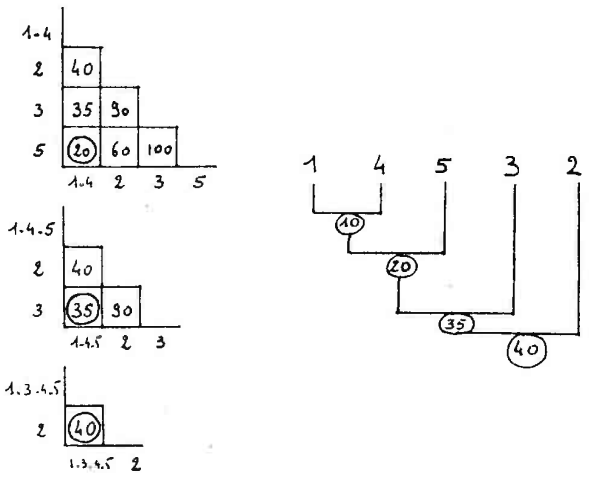


Figure B28b: Classification avec UMI

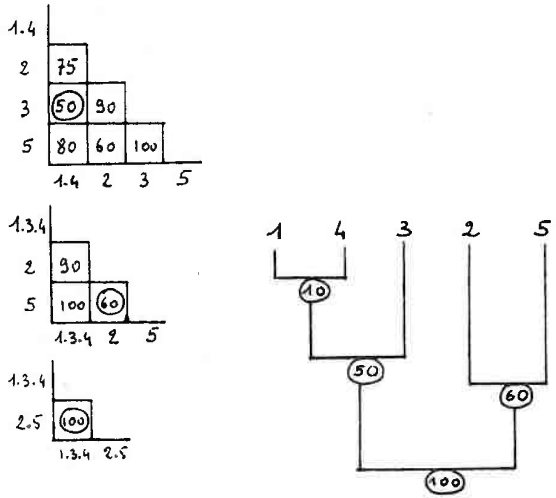


Figure B28c: Classification avec UMS

TABLEAU DE CORRESPONDANCE DES POINTS

No (Clustering)	No locution	Type	Locuteur
1	27	6	3
2	29	1	3
3	33	3	3
4	75	6	7
5	77	1	7
6	81	3	7
7	99	6	9
8	101	1	9
9	105	3	9

:	1	2	3	4	5	6	7	8	9
1:	0	167	228	250	230	238	198	244	327
2:	167	0	158	336	105	134	137	121	237
3:	228	158	0	298	196	94	160	193	145
4:	250	336	298	0	333	252	260	318	266
5:	230	105	196	333	0	161	124	93	231
6:	238	134	94	252	161	0	142	173	117
7:	198	137	160	260	124	142	0	169	224
8:	244	121	193	318	93	173	169	0	213
9:	327	237	145	266	231	117	224	213	0

*** ANALYSE FAITE SELON LE MODE:1 ***
ULTRA-METRIQUE INFERIEURE

NOEUD	INDICE	AINE	BENJAMIN	POIDS
10	93	5	8	2
11	94	3	6	2
12	105	2	10	3
13	117	9	11	3
14	124	7	12	4
15	134	13	14	7
16	167	1	15	8
17	250	4	16	9

:	1	2	3	4	9	6	7	10
1:	0	167	228	250	327	238	198	230
2:	167	0	158	336	237	134	137	105
3:	228	158	0	298	145	94	160	193
4:	250	336	298	0	266	252	260	318
9:	327	237	145	266	0	117	224	213
6:	238	134	94	252	117	0	142	161
7:	198	137	160	260	224	142	0	124
10:	230	105	193	318	213	161	124	0

:	1	2	10	4	9	11	7
1:	0	167	230	250	327	228	198
2:	167	0	105	336	237	134	137
10:	230	105	0	318	213	161	124
4:	250	336	318	0	266	252	260
9:	327	237	213	266	0	117	224
11:	228	134	161	252	117	0	142
7:	198	137	124	260	224	142	0

:	1	7	12	4	9	11
1:	0	198	167	250	327	228
7:	198	0	124	260	224	142
12:	167	124	0	318	213	134
4:	250	260	318	0	266	252
9:	327	224	213	266	0	117
11:	228	142	134	252	117	0

:	1	7	12	4	13
1:	0	198	167	250	228
7:	198	0	124	260	142
12:	167	124	0	318	134
4:	250	260	318	0	252
13:	228	142	134	252	0

:	1	13	14	4
1:	0	228	167	250
13:	228	0	134	252
14:	167	134	0	260
4:	250	252	260	0

:	1	4	15
1:	0	250	167
4:	250	0	252
15:	167	252	0

:	16	4
16:	0	250
4:	250	0

:	17
17:	0

Figure B29a: Evolution de la matrice des distances lors de la classification de 9 locutions en 3 types avec UMI

TABLEAU DE CORRESPONDANCE DES POINTS

No (Clustering)	No locution	Type	Locuteur
1	27	6	3
2	29	1	3
3	33	3	3
4	75	6	7
5	77	1	7
6	81	3	7
7	99	6	9
8	101	1	9
9	105	3	9

:	1	2	3	4	5	6	7	8	9
1:	0	167	228	250	230	238	198	244	327
2:	167	0	158	336	105	134	137	121	237
3:	228	158	0	298	196	94	160	193	145
4:	250	336	298	0	333	252	260	318	266
5:	230	105	196	333	0	161	124	93	231
6:	238	134	94	252	161	0	142	173	117
7:	198	137	160	260	124	142	0	169	224
8:	244	121	193	318	[93]	173	169	0	213
9:	327	237	145	266	231	117	224	213	0

*** ANALYSE FAITE SELON LE MODE:2 ***
ULTRA-METRIQUE SUPERIEURE

NOEUD	INDICE	AINE	BENJAMIN	POIDS
10	93	5	8	2

:	1	2	3	4	9	6	7	10
1:	0	167	228	250	327	238	198	244
2:	167	0	158	336	237	134	137	121
3:	228	158	0	298	145	94	160	196
4:	250	336	298	0	266	252	260	333
9:	327	237	145	266	0	117	224	231
6:	238	134	[94]	252	117	0	142	173
7:	198	137	160	260	224	142	0	169
10:	244	121	196	333	231	173	169	0

NOEUD	INDICE	AINE	BENJAMIN	POIDS
11	94	3	6	2

:	1	2	10	4	9	11	7
1:	0	167	244	250	327	238	198
2:	167	0	121	336	237	158	137
10:	244	[121]	0	333	231	196	169
4:	250	336	333	0	266	298	260
9:	327	237	231	266	0	145	224
11:	238	158	196	298	145	0	160
7:	198	137	169	260	224	160	0

NOEUD	INDICE	AINE	BENJAMIN	POIDS
12	121	2	10	3

:	1	7	12	4	9	11
1:	0	198	244	250	327	238
7:	198	0	169	260	224	160
12:	244	169	0	336	237	196
4:	250	260	336	0	266	298
9:	327	224	237	266	0	145
11:	238	160	196	298	[145]	0

NOEUD	INDICE	AINE	BENJAMIN	POIDS
13	145	9	11	3

:	1	7	12	4	13
1:	0	198	244	250	327
7:	198	0	169	260	224
12:	244	[169]	0	336	237
4:	250	260	336	0	298
13:	327	224	237	298	0

NOEUD	INDICE	AINE	BENJAMIN	POIDS
14	169	7	12	4

:	1	13	14	4
1:	0	327	244	250
13:	327	0	237	298
14:	244	[237]	0	336
4:	250	298	336	0

NOEUD	INDICE	AINE	BENJAMIN	POIDS
15	237	13	14	7

:	1	4	15
1:	0	250	327
4:	[250]	0	336
15:	327	336	0

NOEUD	INDICE	AINE	BENJAMIN	POIDS
16	250	1	4	2

:	15	16
15:	0	336
16:	[336]	0

NOEUD	INDICE	AINE	BENJAMIN	POIDS
17	336	15	16	9

:	17
17:	0

Figure B29b: Evolution de la matrice des distances lors de la classification de 9 locutions en 3 types avec UMS

Résultats

Le résultat le plus évident et qui a déjà été constaté dans d'autres domaines ([Chandon 80]), ([Pister 84]) est le chainage dû à l'ultramétrique inférieure. La définition de cette mesure entraîne l'agrégation des locutions de proche en proche jusqu'à la constitution de classes très ''allongées'' : les chaînes. Un exemple réduit en est donné figure B28 avec cinq points, l'arbre associé à l'UMI présente une structure hiérarchique en cascade typique du chaînage. (Ce qui n'est pas le cas de l'UMS appliquée au même corpus) (cf. aussi figure B30a et B30b). Ces arbres en ''cascade'' présentent deux inconvénients majeurs :

- lors de leur segmentation en classes (cf. paragraphe 635) ; ils donnent de nombreux points isolés.
- ils forment des classes ''allongées'' c'est-à-dire que dans une même classe peuvent être agrégés deux points très éloignés pourvu qu'il y ait entre eux une chaîne continue. Ces types de classes ne sont pas compatibles avec la notion de représentant et de rayon choisie (cf. paragraphe 64) qui ne rend bien compte que de classes ''sphériques''.

On voit sur les deux figures B30a et B30b que la structure en cascade dépend non seulement des caractéristiques du corpus mais aussi de la mesure utilisée.

Si l'on compare le nombre de classes obtenues dans les mêmes conditions et sur le même corpus (les chiffres français) selon les quatres critères d'agrégations on obtient les résultats présentés sur la figure B31a.

On constate que l'ultramétrique inférieure se révèle inadaptée au projet, en effet, le chaînage entraîne un émiettement important des classes lors de l'étape de segmentation (cf paragraphe suivant); elle a donc été abandonnée et les tests de reconnaissance ultérieurs ne concernent que les trois autres critères. Les résultats obtenus (cf. figure B31b et B31c) montrent que ces trois critères sont eux, bien adaptés au problème et fournissent des taux de reconnaissance satisfaisants et assez proches les uns des autres. Le critère de la variance apporte cependant un avantage assez net (quatre points en première proposition) qui s'est de plus révélé remarquablement constant au cours des différents tests effectués.

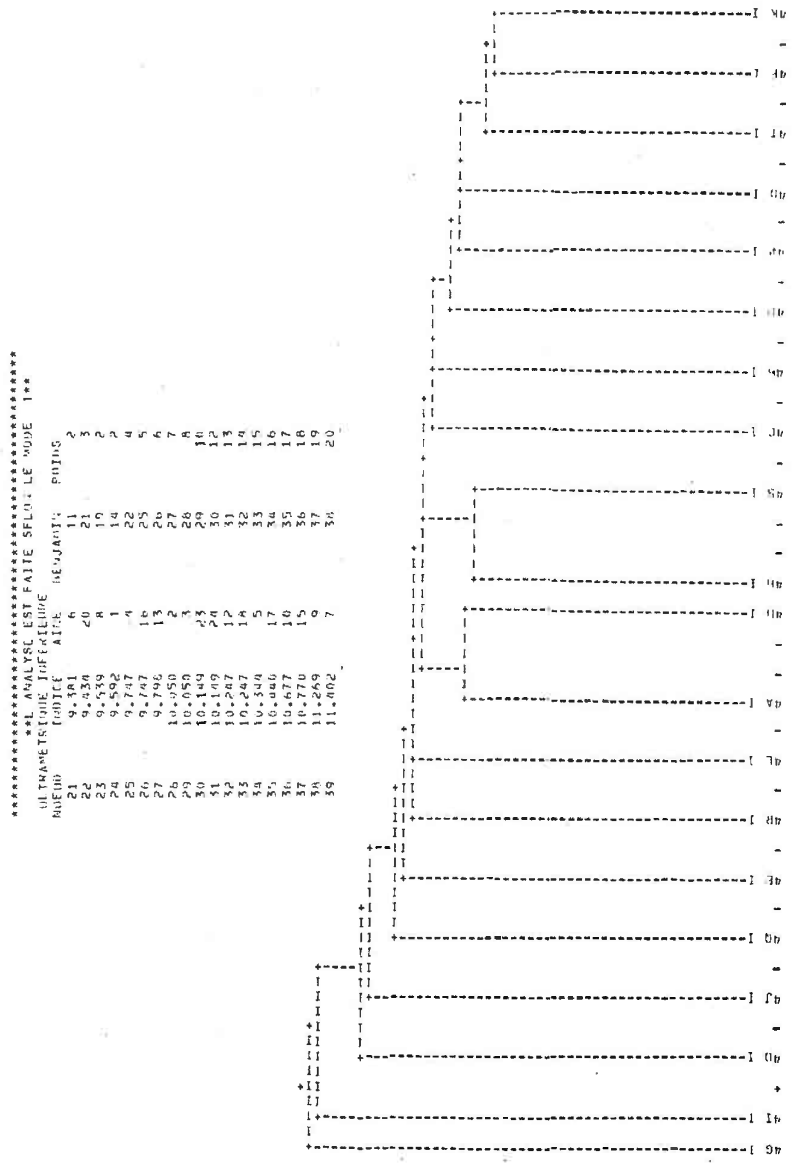


Figure B30a: Tableau et arbre hiérarchique correspondant à la classification de 20 occurrences du mot "quatre " avec UMI. On constate la formation de chaines caractéristique de cette métrique.

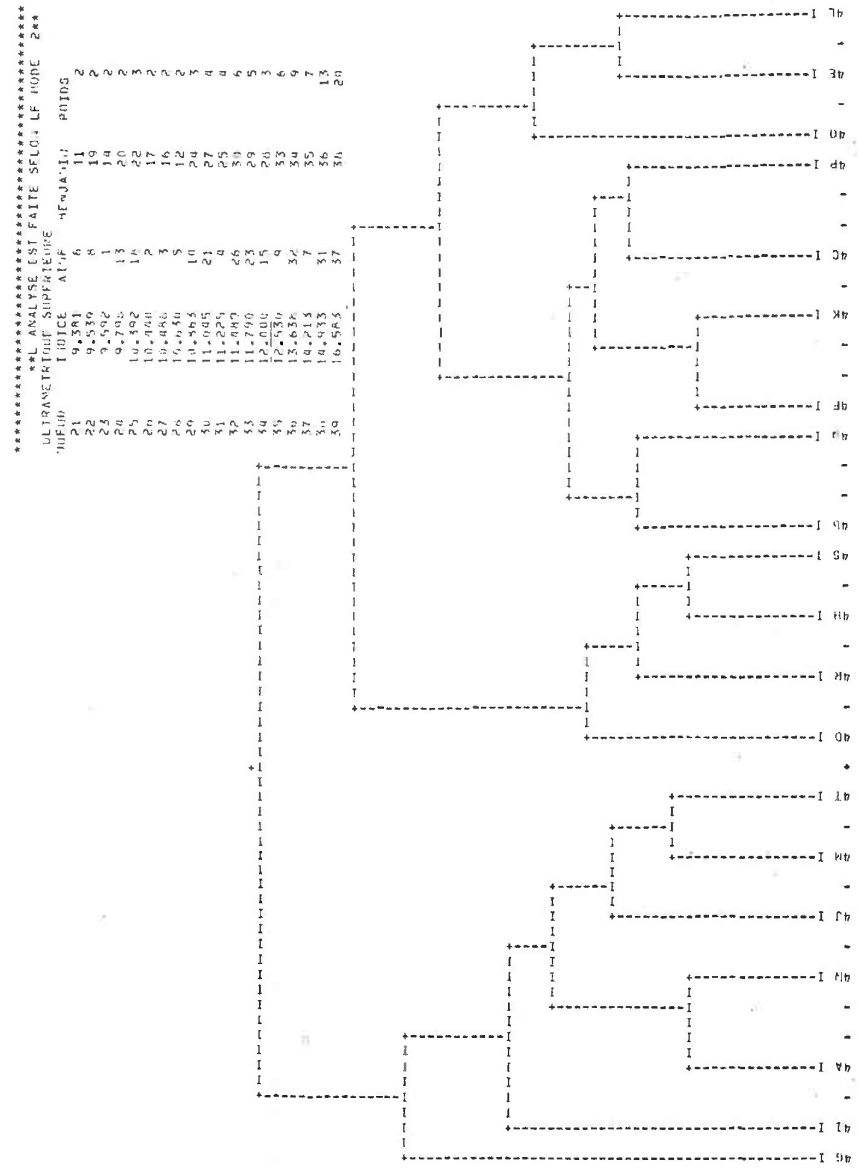


Figure B30b: Tableau et arbre hiérarchique correspondant à la classification de 20 occurrences du mot "quatre " avec UMS. (la formation de chaines est ici beaucoup moins fréquente)

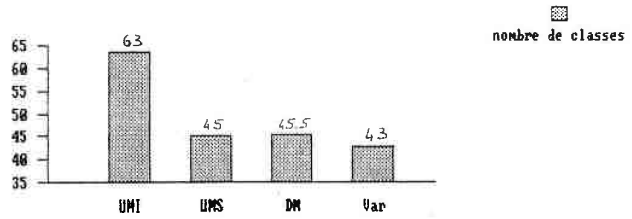


Figure B31a: Comparaison du nombre de classes obtenu pour chaque métrique

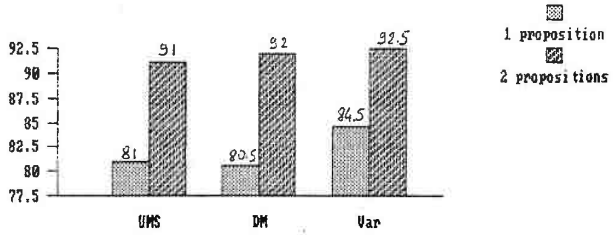


Figure B31b: Comparaison des taux de reconnaissance pour la classification globale

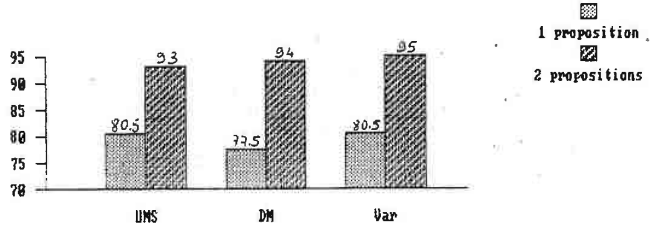


Figure B31c: Comparaison des taux de reconnaissance pour la classification par type

635. Stratégie de segmentation de l'arbre

a) Principe

Cette phase ne concerne que la méthode de classification ascendante hiérarchique. Nous avons vu dans les paragraphes précédents comment on obtenait un arbre hiérarchique de locutions. Le but de cette étape est d'obtenir des classes c'est-à-dire des ensembles de locutions vérifiant le mieux possible les caractères suivants :

(a) - densité : faibles distances intra-classe.

(b) - différenciation : fortes distances inter-classe.

(c) - homogénéité : forte proportion de locutions du même type dans une classe.

(d) - exclusion : pas d'intersection de classes.

(e) - totalité : l'ensemble des classes recouvre (ou presque) la totalité du corpus.

Ces classes sont obtenues en segmentant l'arbre hiérarchique, elles sont constituées des feuilles de chaque branche coupée comme illustré figure B32.

- Les critères a et b découlent des caractéristiques des algorithmes de classification (pertinence des classes).
- Le critère c est dû au fait que de chaque classe va être extraite une seule référence représentative d'un type. Ce critère n'est évidemment pas pertinent pour la classification par type.
- Le critère d est obligatoirement satisfait du fait de la structure arborescente.
- Le critère e impose que les classes obtenues rendent compte de la totalité de l'information contenue dans le corpus d'apprentissage. On acceptera toutefois quelques exceptions dues à des points isolés. (cf. paragraphe 65)
- Les critères d et e font de l'ensemble des classes une quasi-partition du corpus d'apprentissage.

Les stratégies de segmentation de l'arbre s'ordonnent suivant deux dimensions :

- Le choix d'un mode de segmentation concernant le niveau,
- Le choix d'un critère de guidage de cette segmentation.

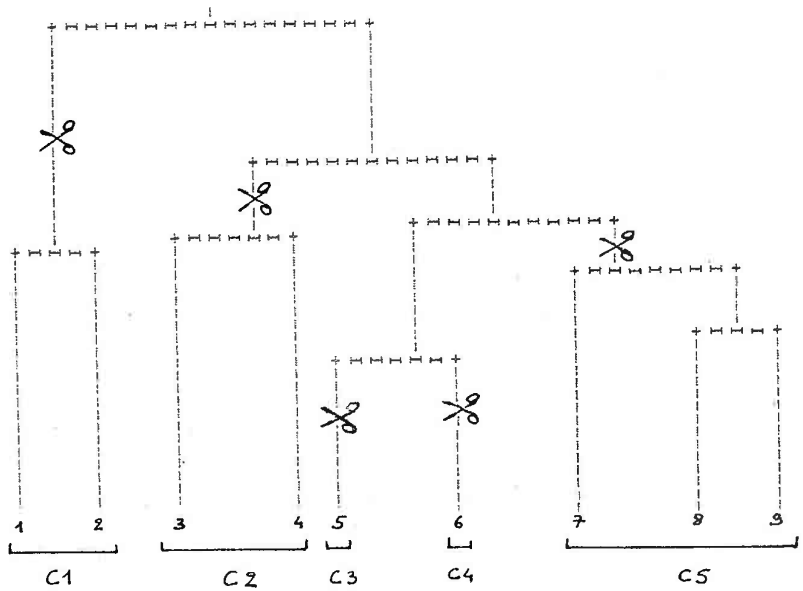


Figure B32: 5 classes obtenues par la segmentation d'un arbre de 9 points en 5 branches

b) Mode de segmentation

* Segmentation à niveau fixe.

Puisqu'il s'agit de "casser" les branches de l'arbre de façon à obtenir une partition de l'ensemble des locutions, l'idée la plus simple consiste à sélectionner une des partitions Pi de la hiérarchie (cf. Partie A ,paragraphe 531) ; ce qui revient à segmenter l'arbre horizontalement à niveau fixe, c'est-à-dire à choisir un indice d'agrégation en dessous duquel les branches forment des classes. (cf figure B33a).

* Segmentation à niveau variable

On s'autorise cette fois à couper les branches différentes à des hauteurs différentes. On obtient encore une partition du corpus (puisqu'une même branche ne peut pas être coupée à deux niveaux différents) mais les classes qui la forment possèdent des indices d'agrégation éventuellement très variables. (cf. figure B33b)

c) Critère de guidage

Quel que soit le mode utilisé, il reste encore à déterminer à quelle(s) hauteur(s) les branches de l'arbre doivent être coupées.

Pour cela, l'algorithme peut être guidé par plusieurs critères :

* Critère de type

Utilisable seulement dans le cas d'une classification globale. La segmentation est guidée par l'homogénéité des classes qui en sont issues c'est-à-dire qu'une branche est coupée lorsque toutes ses feuilles-locutions sont de même type (ou du moins une forte proportion).

* Critères numériques

Utilisable dans le cas d'une classification par type, où le précédent critère ne peut pas s'appliquer. La segmentation est guidée par un critère numérique sur les classes. Ce nombre est fixé à priori, c'est un paramètre du système.

- Segmentation guidée par le nombre total de classes.
- Segmentation guidée par le nombre maximum de points autorisés pour une classe.

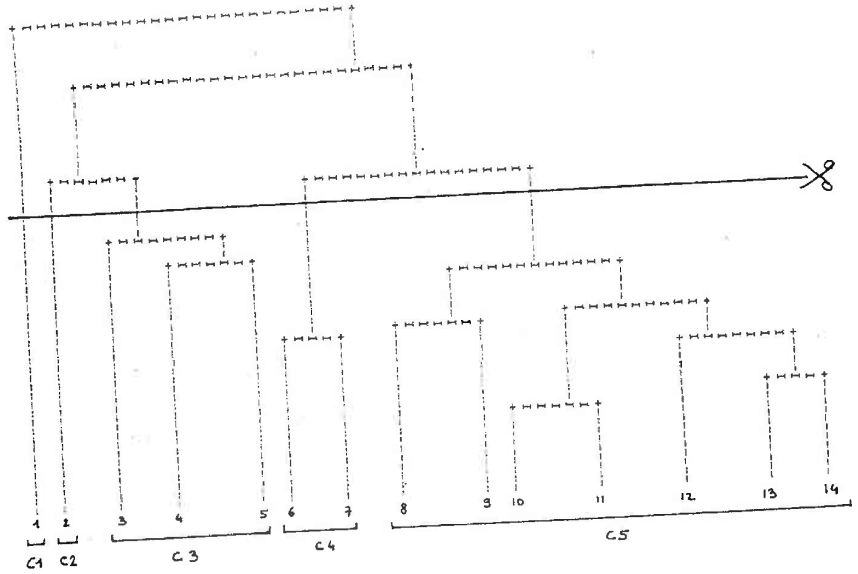


Figure B33a: Segmentation à niveau fixe (5 classes dont 2 singletons)

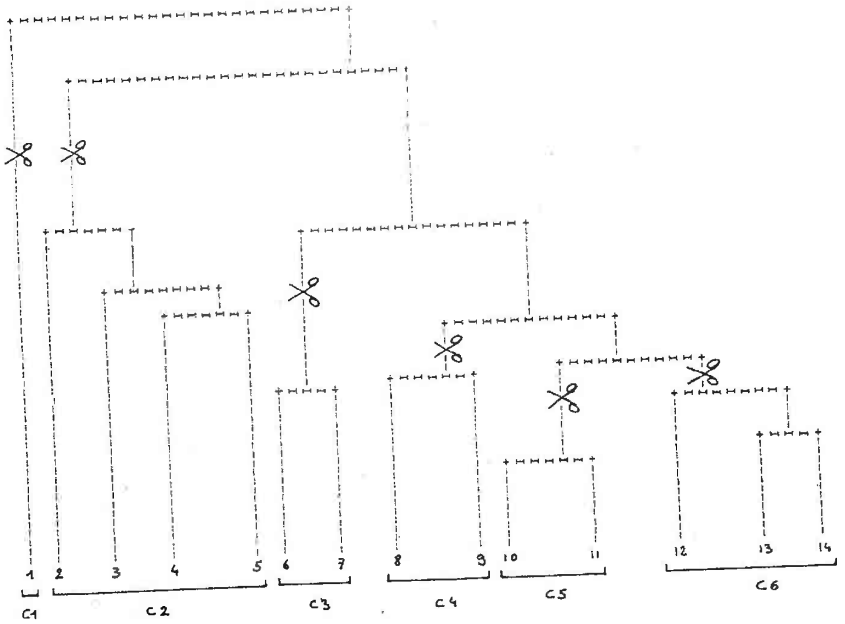


Figure B33b: Segmentation à niveau variable (6 classes dont 1 singleton)

Ces différentes stratégies de segmentation donnent lieu en principe à six combinaisons possibles:

Mode de segmentation	Critère de guidage	Type de classif.
à niveau fixe	Nombre total de classes	Par type
à niveau fixe	Nb de points max/classe	Par type
à niveau variable	Nombre total de classes	Par type
à niveau variable	Nb de points max/classe	Par type
à niveau variable	Type des locutions	Globale

Nous avons vu au chapitre 5 que certains types de locutions étaient denses et d'autres lâches. Imaginons le cas de trois types (1 lâche et 2 denses) répartis comme indiqué sur la figure B34.
L'arbre hiérarchique associé sera celui de la figure B35.

Une segmentation à niveau constant rend difficilement compte de cette répartition, en effet une coupure au niveau A confond les types 1 et 2 dans la même classe et une coupure au niveau B qui sépare correctement ces deux types découpe le type 1 en un grand nombre de points isolés.
Cette considération nous a fait éliminer a priori des stratégies à tester les combinaisons FT et FM.

Le choix FC reste cependant possible, car, dans le cas d'un classement par type, une segmentation au niveau A est moins gênante (cela se traduit par le regroupement dans une même classe de deux prononciations légèrement différentes d'un même mot).

D'autre part le nombre de classes ne peut pas être un critère pour une segmentation à niveau variable, ce qui élimine la combinaison VC.

Seules restent donc trois combinaisons :
Pour la classification par type :

- Segmentation à niveau fixe guidée par le nombre total de classes (FC).
- Segmentation à niveau variable guidée par le nombre de points maximum par classe (VM).

Pour la classification globale :

- Segmentation à niveau variable guidée par le type des locutions-feuilles (VT)

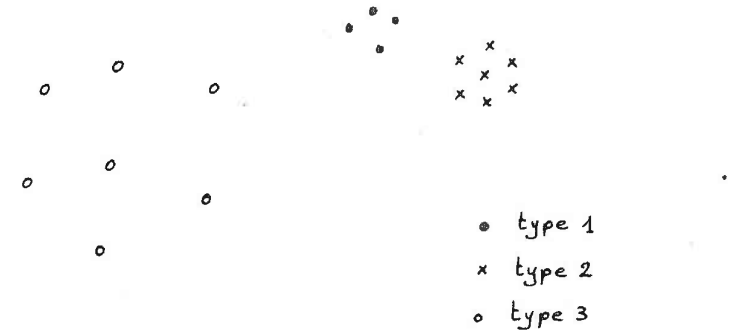


Figure B34: Exemple schématique de 18 locutions réparties en 3 types

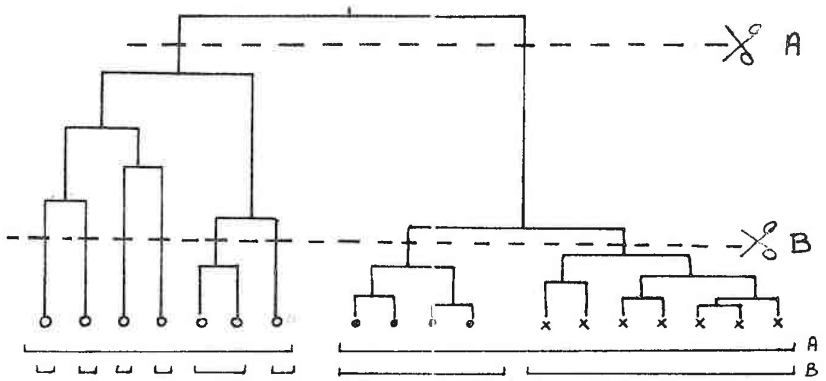


Figure B35: L'arbre hiérarchique correspondant
et deux segmentations possibles

d) Algorithmes

1- Segmentation à niveau fixe

+ Cas général: Algorithme récursif non guidé(Le seuil S est fixé à l'avance), initié par l'instruction SEG(Tr) où Tr est le tronc de l'arbre (c'est à dire le noeud 2N+1).

SEG(Noeud)

Si indice(Noeud) ≤ S alors les feuilles issues de Noeud forment une classe qu'on mémorise.
Sinon SEG(Ainé(Noeud)) ; SEG(Benj(Noeud))

Remarque: Les feuilles de l'arbre sont les locutions du corpus de départ, elles ont par convention un indice nul.

+ Segmentation à niveau fixe guidée par le nombre total de classes (classification par type). Algorithme non-récursif où N est le nombre de locutions, C le nombre de classes désiré.
Précondition: 1 ≤ C ≤ N

SEGFC
Si C=1 alors les feuilles du noeud 2N-1 forment la classe cherchée
Sinon pour i de 2N+1-C à 2N-1 répéter
Si Ainé(i) < 2N+1-C alors les feuilles de Ainé(i) forment une classe qui est mémorisée.
Si Benj(i) < 2N+1-C alors les feuilles de Benj(i) forment une classe qui est mémorisée.

Voir exemple sur la figure B36.

Avantages :
Etant directement lié au temps de reconnaissance , le nombre de classes désiré par types de mot est un paramètre assez facile à déterminer a priori.

Inconvénients :
Peu de contrôle sur la pertinence des classes (voir exemple figures B34, B35).
Rigidité de la classification car sauf décision humaine tous les

Matrice des distances de 15 mots

TABLEAU DE CORRESPONDANCE DES POINTS

No (Clustering)	No locution	Type	Locuteur
1	4	4	1
2	16	4	2
3	28	4	3
4	40	4	4
5	52	4	5
6	64	4	6
7	76	4	7
8	88	4	8
9	100	4	9
10	112	4	10
11	124	4	11
12	136	4	12
13	148	4	13
14	160	4	14
15	172	4	15

*** ANALYSE FAITE SELON LE MODE:2 ***
ULTRA-METRIQUE SUPERIEURE

NOEUD	INDICE	AINE	BENJAMIN	POIDS
16	95.	6	11	2
17	100.	1	14	2
18	107.	5	15	2
19	114.	3	4	2
20	118.	8	12	2
21	118.	2	16	3
22	134.	13	19	3
23	147.	17	18	4
24	153.	20	21	5
25	162.	22	23	7
26	164.	9	10	2
27	170.	25	26	9
28	180.	7	24	6
29	200.	27	28	15

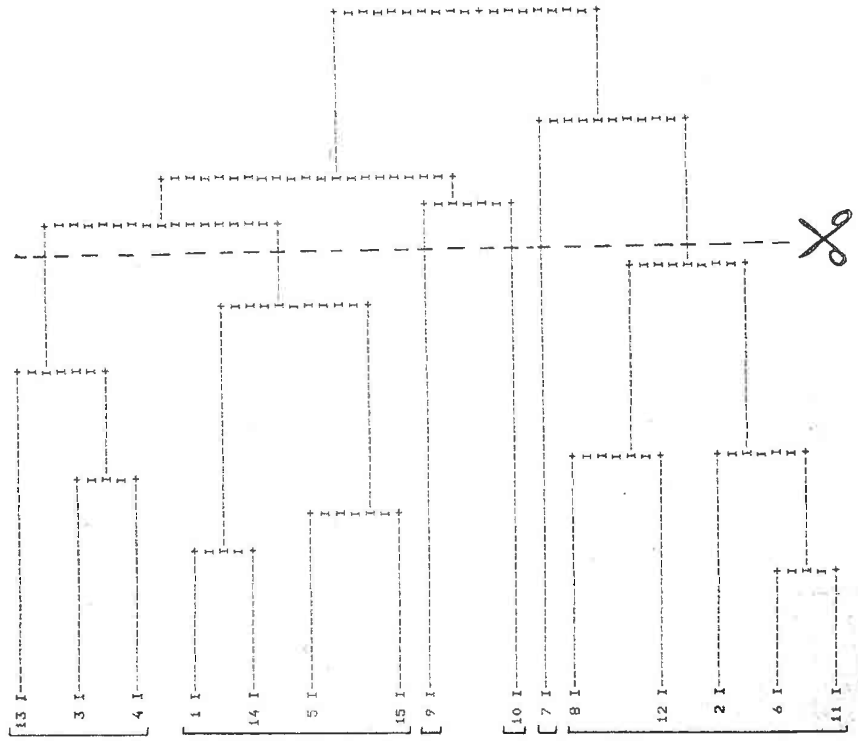


Figure B36: Segmentation à niveau fixe guidée par le nombre de classes (ici 6), effectuée sur 15 énonciations du mot "quatre "

types auront le même nombre de représentants ce qui ne correspond pas à la diversité du corpus. De plus le nombre total ne peut varier que par bonds importants. (Dans notre application pourtant restreinte, si on passe de quatre à cinq classes par type de mot, le nombre de références passera de quarante à cinquante).

2- Segmentation à niveau variable guidée par le nombre de points maximum autorisés par classe (classification par type).
Algorithme récursif descendant à des profondeurs différentes dans chaque branche. Il est initié par l'instruction SCIE(Tr) où Tr est le tronc de l'arbre, c'est-à-dire le noeud de la dernière agrégation.

SCIE (Noeud)

* Lister toutes les feuilles issues de Noeud ----> Flles

* Si card (Flles) > Maxclu

  alors SCIE (Ainé (Noeud))
  SCIE (Benj (Noeud))

  sinon Flles forme une classe qu'on mémorise

Ainé et Benj sont les deux fils du noeud : Noeud.
Maxclu est le paramètre : nombre de locutions maximum autorisé dans une classe.
Voir exemple figure B37

Avantages

Bonne représentativité des références même dans le cas de corpus hétérogène.

Nombre de représentants adapté à la diversité des prononciations d'un type.

Inconvénients

Pas de contrôle direct sur le nombre de classes. Un contrôle indirect est cependant possible par le biais du nombre de points maximum par

Matrice lue sur fichier :MQUAT15
Matrice des distances de 15 mots

TABLEAU DE CORRESPONDANCE DES POINTS

No (Clustering)	No locution	Type	Locuteur
1	4	4	1
2	16	4	2
3	28	4	3
4	40	4	4
5	52	4	5
6	64	4	6
7	76	4	7
8	88	4	8
9	100	4	9
10	112	4	10
11	124	4	11
12	136	4	12
13	148	4	13
14	160	4	14
15	172	4	15

*** ANALYSE FAITE SELON LE MODE:2 ***
ULTRA-METRIQUE SUPERIEURE

NOEUD	INDICE	AINE	BENJAMIN	POIDS
16	95.	6	11	2
17	100.	1	14	2
18	107.	5	15	2
19	114.	3	4	2
20	118.	8	12	2
21	118.	2	16	3
22	134.	13	19	3
23	147.	17	18	4
24	153.	20	21	5
25	162.	22	23	7
26	164.	9	10	2
27	170.	25	26	9
28	180.	7	24	6
29	200.	27	28	15

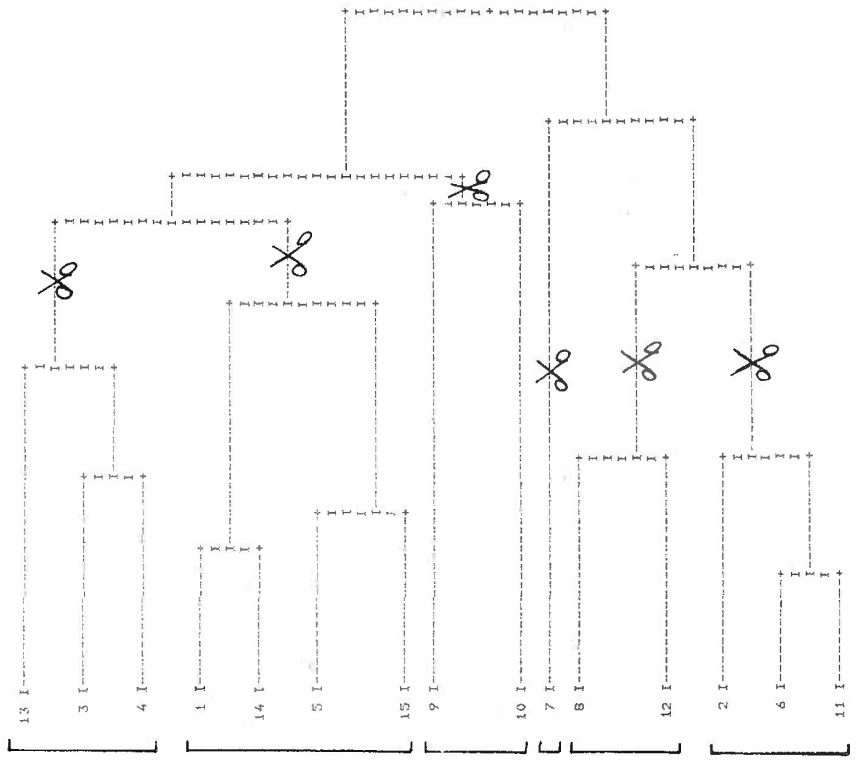


Figure B37: Segmentation à niveau variable guidée par le nombre de points maximum par classe (ici 4), effectuée sur 15 énonciations du mot "quatre "

classe. (cf. résultats), mais il est difficile de fixer correctement la valeur de ce paramètre a priori.

3- Segmentation à niveau variable guidée par le type des feuilles (classification globale). Algorithme récursif initié par l'instruction SCIE(TR) où TR est le tronc de l'arbre.

SCIE (Noeud)

- lister toutes les feuuilles issues de Noeud ----› Flles
- Si Flles ne contient que des locutions de même type
  alors Flles forme une classe
  sinon
  SCIE (Ainé (Noeud))
  SCIE (Benj (Noeud))

où Ainé et Benj sont les deux fils de Noeud.
Voir exemple figure B38

Avantages

Rend compte aussi fidèlement qu'on le désire de l'homogénéité des classes selon leur type,

Nombre de représentants d'un type adapté non seulement à la diversité des prononciations mais aussi et surtout à la proximité des autres types.

Inconvénients

Pas de contrôle direct sur le nombre de représentants.

Emiettement des classes en cas de types très empiétants.

e) Taux de recouvrement

Afin de réduire ce dernier inconvénient, nous avons introduit dans l'algorithme de segmentation un coefficient d'homogénéité (coefhomo) autorisant une légère proportion d'''intrus'' dans une classe (c'est-à-dire de locutions d'un type différent). La conditionnelle devient alors :

Si Flles contient au moins une proportion de Coefhomo de locutions du même type

alors Flles - {les locutions d'autres types} forme une classe.

Ce pourcentage Coefhomo est un paramètre du système qui permet, par l'intermédiaire d'un certain taux de recouvrement autorisé d'agir sur le nombre de classes représentées lors de la reconnaissance (cf figure B39).
N.B : Les ''intrus'' sont purement et simplement éliminés du corpus, ce qui explique la restriction émise lors du critère de totalité (cf. paragraphe 635 a).

Remarque 1 : Dans aucun des trois algorithmes envisagés l'indice d'agrégation de la classe n'est mémorisé pour l'étape de reconnaissance; en effet cette information sur la densité de la classe est retrouvée ultérieurement par le biais du ''rayon d'influence ''(cf, paragraphe 64).

Remarque 2 : Une autre méthode (C.S.A) utilisant l'arbre hiérarchique a été testée ; dans ce cas l'arbre n'est pas segmenté mais étiqueté en vue d'accélérer l'étape de reconnaissance. Etant assez différente des autres méthodes (classification et représentants sont remplacés par une structuration arborescente des locutions d'apprentissage) elles sera présentée à part au paragraphe 67.

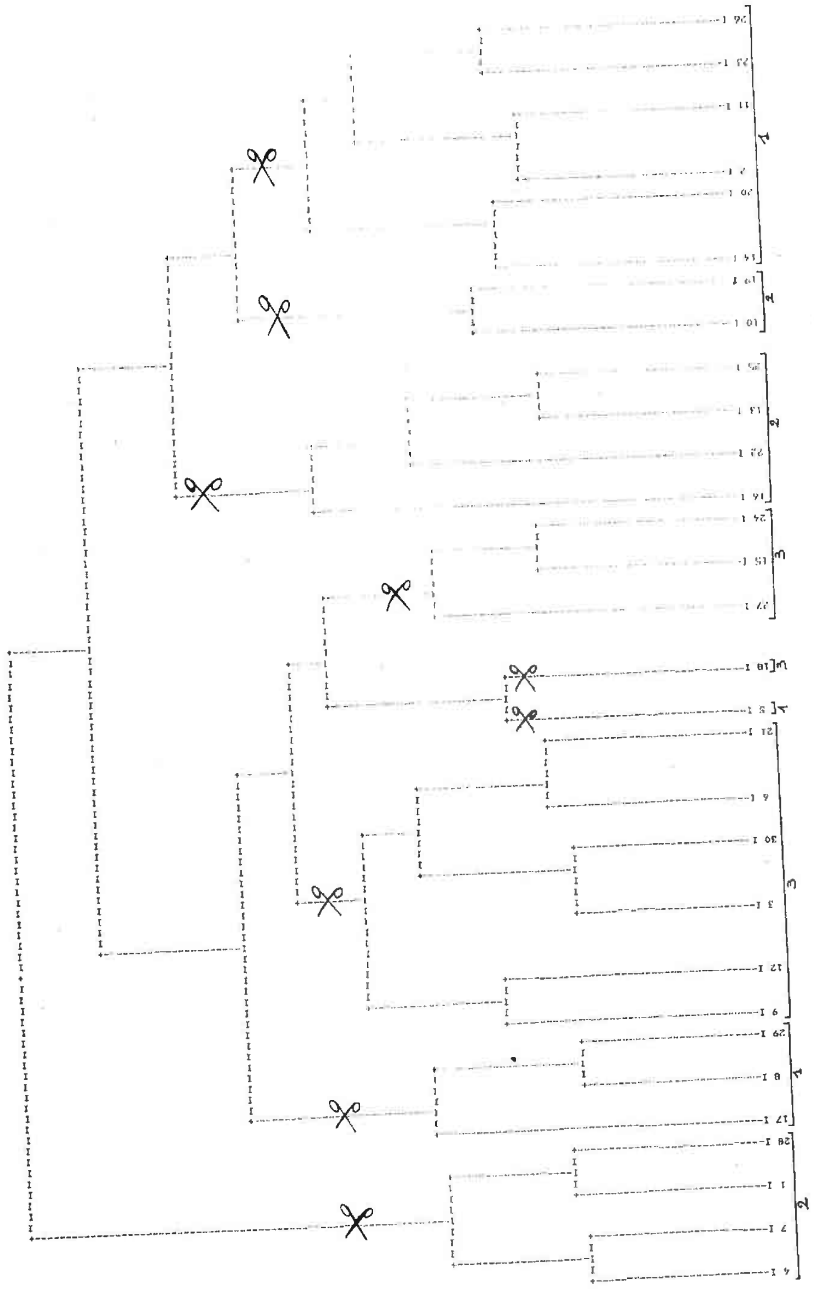


Figure B38a: Segmentation à niveau variable guidée par le type effectuée sur 30 mots répartis en 3 types

Matrice des distances de 30 mots

TABLEAU DE CORRESPONDANCE DES POINTS

No (Clustering)	No locution	Type	Locuteur
1	1	2	1
2	5	1	1
3	9	3	1
4	13	2	2
5	17	1	2
6	21	3	2
7	25	2	3
8	29	1	3
9	33	3	3
10	37	2	4
11	41	1	4
12	45	3	4
13	49	2	5
14	53	1	5
15	57	3	5
16	61	2	6
17	65	1	6
18	69	3	6
19	73	2	7
20	77	1	7
21	81	3	7
22	85	2	8
23	89	1	8
24	93	3	8
25	97	2	9
26	101	1	9
27	105	3	9
28	109	1	10
29	113	3	10
30	117	2	10

*** CLASSE FAITE SELON LE MODELE ***
NOMBRE D'UNE CLASSE

NOEUD	TAILLE	GAIN	DELIGNATO	POIDS
31	2.	0	2	2
32	2.	0	16	2
33	2.	0	8	2
34	2.	0	4	2
35	30.	13	26	2
36	28.	15	24	2
37	28.	6	21	2
38	26.	2	11	2
39	26.	14	20	2
40	26.	5	18	2
41	27.	23	26	2
42	28.	9	12	2
43	29.	10	19	2
44	27.	27	36	3
45	27.	34	34	4
46	29.	17	31	3
47	29.	22	35	3
48	31.	32	37	4
49	36.	38	41	4
50	38.	42	48	6
51	43.	40	44	5
52	44.	39	49	6
53	45.	16	47	4
54	69.	50	51	11
55	74.	43	52	8
56	77.	46	54	14
57	85.	53	55	12
58	99.	56	57	26
59	111.	45	58	30

RESULTATS DU CLUSTERING

Coefficient d'homogeneite pour un cluster : 1.00

CLUSTER DE TYPE 2:

Numerotation clustering: 4 7 1 28
Numerotation ds TLOCUT : 13 25 1 109
Apres elim. des P.I. : 13 25 1 109

CENTRE DU CLUSTER: 4(13)
RAYON: 124.00

CLUSTER DE TYPE 1:

Numerotation clustering: 17 8 29
Numerotation ds TLOCUT : 65 29 113
Apres elim. des P.I. : 65 29 113

CENTRE DU CLUSTER: 8(29)
RAYON: 155.00

CLUSTER DE TYPE 3:

Numerotation clustering: 9 12 3 30 6 21
Numerotation ds TLOCUT : 33 45 9 117 21 81
Apres elim. des P.I. : 33 45 9 117 21 81

CENTRE DU CLUSTER: 6(21)
RAYON: 143.00

CLUSTER DE TYPE 1:

Numerotation clustering: 5
Numerotation ds TLOCUT : 17
Apres elim. des P.I. : 17

CENTRE DU CLUSTER: 5(17)
RAYON: 0.00

CLUSTER DE TYPE 3:

Numerotation clustering: 18
Numerotation ds TLOCUT : 69
Apres elim. des P.I. : 69

CENTRE DU CLUSTER: 18(69)
RAYON: 0.00

CLUSTER DE TYPE 3:

Numerotation clustering: 27 15 24
Numerotation ds TLOCUT : 105 57 93
Apres elim. des P.I. : 105 57 93

CENTRE DU CLUSTER: 15(57)
RAYON: 135.00

CLUSTER DE TYPE 2:

Numerotation clustering: 16 22 13 25
Numerotation ds TLOCUT : 61 85 49 97
Apres elim. des P.I. : 61 85 49 97

CENTRE DU CLUSTER: 22(85)
RAYON: 199.00

CLUSTER DE TYPE 2:

Numerotation clustering: 10 19
Numerotation ds TLOCUT : 37 73
Apres elim. des P.I. : 37 73

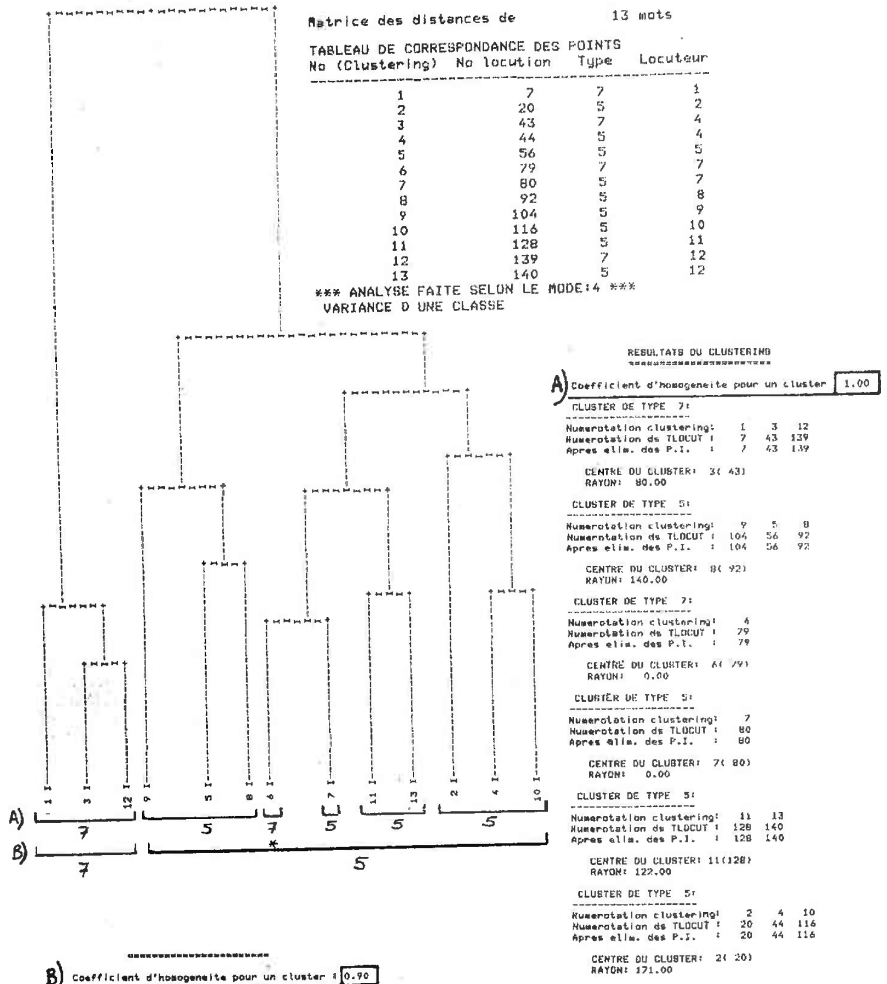
CENTRE DU CLUSTER: 10(37)
RAYON: 156.00

CLUSTER DE TYPE 1:

Numerotation clustering: 14 20 2 11 23 26
Numerotation ds TLOCUT : 53 77 5 41 89 101
Apres elim. des P.I. : 53 77 5 41 89 101

CENTRE DU CLUSTER: 14(53)
RAYON: 163.00

Figure B38b: Classification correspondant à l'arbre précédent (Coefhomo = 1) . Noter qu'un Coefhomo = 0.9 entrainerait la disparition de la locution 5 et la création d'une seule classe pour le mot ''trois''



Matrice des distances de 13 mots

TABLEAU DE CORRESPONDANCE DES POINTS

No (Clustering)	No locution	Type	Locuteur
1	7	7	1
2	20	5	2
3	43	7	4
4	44	5	4
5	56	5	5
6	79	7	7
7	80	5	7
8	92	5	8
9	104	5	9
10	116	5	10
11	128	5	11
12	139	7	12
13	140	5	12

*** ANALYSE FAITE SELON LE MODE:4 ***
VARIANCE D UNE CLASSE

A) 7 5 7 5 5 5

B) 7 5

RESULTATS DU CLUSTERING

A) Coefficient d'homogeneite pour un cluster : 1.00

CLUSTER DE TYPE 7:

Numerotation clustering: 1 3 12
Numerotation ds TLOCUT : 7 43 139
Apres elim. des P.I. : 7 43 139

CENTRE DU CLUSTER: 3(43)
RAYON: 80.00

CLUSTER DE TYPE 5:

Numerotation clustering: 9 5 8
Numerotation ds TLOCUT : 104 56 92
Apres elim. des P.I. : 104 56 92

CENTRE DU CLUSTER: 8(92)
RAYON: 140.00

CLUSTER DE TYPE 7:

Numerotation clustering: 6
Numerotation ds TLOCUT : 79
Apres elim. des P.I. : 79

CENTRE DU CLUSTER: 6(79)
RAYON: 0.00

CLUSTER DE TYPE 5:

Numerotation clustering: 7
Numerotation ds TLOCUT : 80
Apres elim. des P.I. : 80

CENTRE DU CLUSTER: 7(80)
RAYON: 0.00

CLUSTER DE TYPE 5:

Numerotation clustering: 11 13
Numerotation ds TLOCUT : 128 140
Apres elim. des P.I. : 128 140

CENTRE DU CLUSTER: 11(128)
RAYON: 122.00

CLUSTER DE TYPE 5:

Numerotation clustering: 2 4 10
Numerotation ds TLOCUT : 20 44 116
Apres elim. des P.I. : 20 44 116

CENTRE DU CLUSTER: 2(20)
RAYON: 171.00

B) Coefficient d'homogeneite pour un cluster : 0.90

CLUSTER DE TYPE 7:

Numerotation clustering: 1 3 12
Numerotation ds TLOCUT : 7 43 139
Apres elim. des P.I. : 7 43 139

CENTRE DU CLUSTER: 3(43)
RAYON: 80.00

CLUSTER DE TYPE 5:

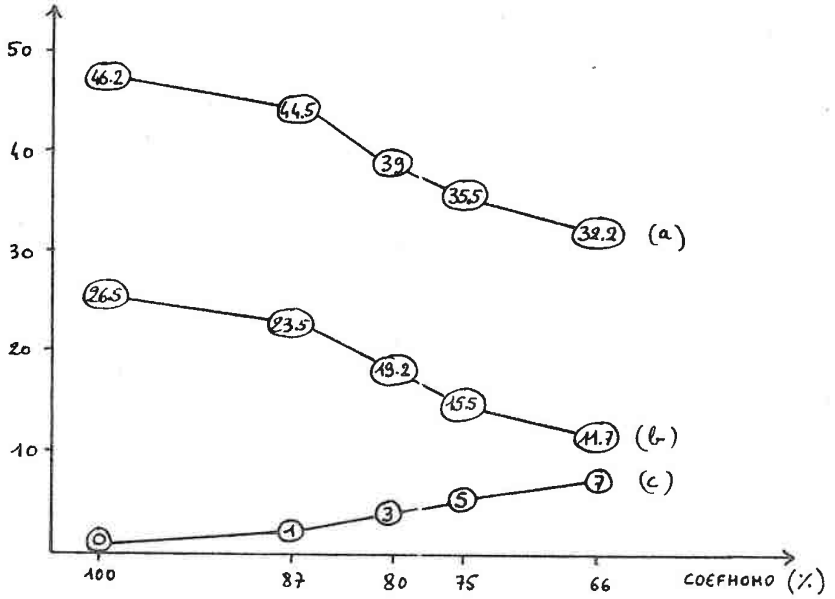
Numerotation clustering: 9 5 8 6* 7 11 13 2 4 10
Numerotation ds TLOCUT : 104 56 92 79* 80 128 140 20 44 116
Apres elim. des P.I. : 104 56 92 80 128 140 20 44 116

CENTRE DU CLUSTER: 4(44)
RAYON: 176.00

Figure B39: Influence du coefficient d'homogénéité sur l'émiettement des classes. Classification guidée par le type de 13 énonciations des mots ''cinq'' et ''sept''

f) Résultats

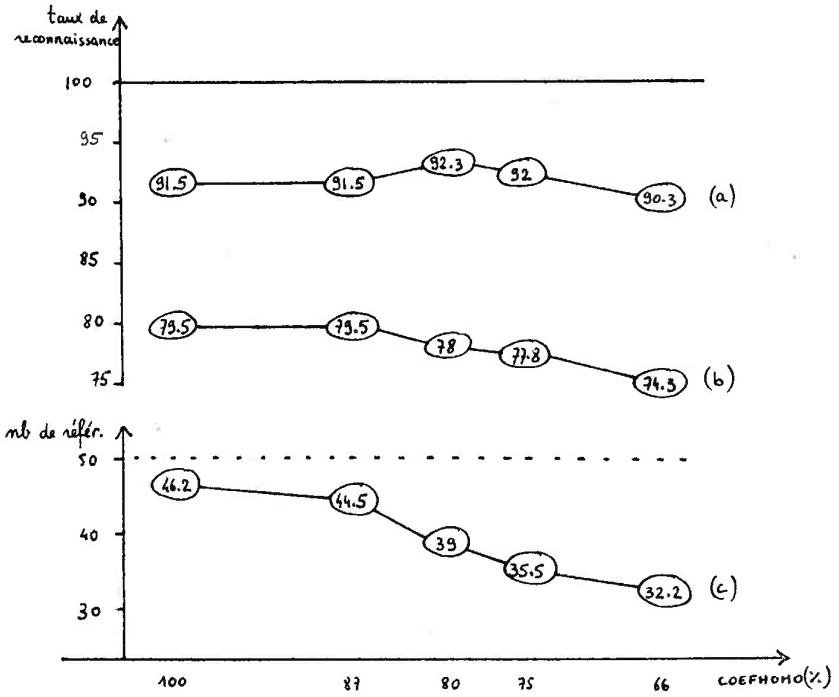
1) Influence du coefficient d'homogénéité sur la répartition des classes



(a) : nombre total de classes représentées
(b) : nombre de classes réduites à un point
(c) : nombre de points supprimés (intrus)

Figure B40: Evolution du nombre de classes en fonction du coefficient de recouvrement COEFHOMO pour 10 types de mots.

2) Influence du coefficient d'homogénéité sur le taux de reconnaissance



(a) : taux de reconnaissance en deux propositions
(b) : taux de reconnaissance en une proposition
(c) : nombre de classes (pour dix types de mots)

Figure B41: Evolution du taux de reconnaissance en fonction du coefficient de recouvrement COEFHOMO.

3) Influence du type de segmentation sur les taux de reconnaissance

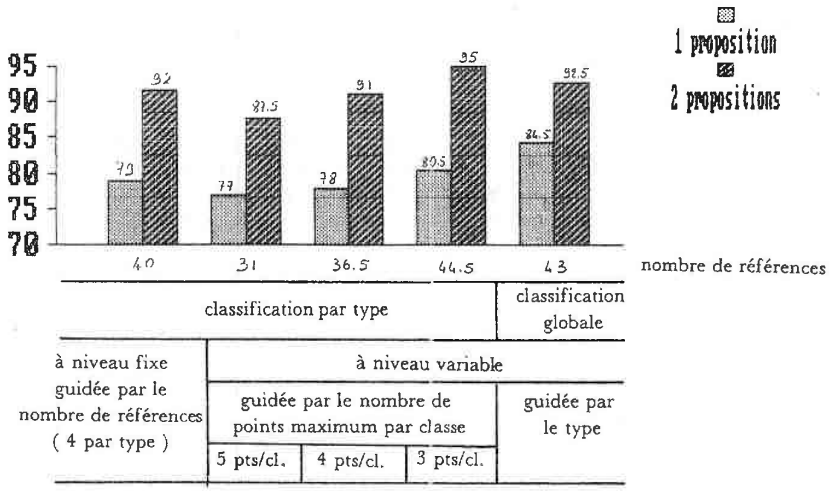


Figure B42: Récapitulatif des différents types de segmentation et influence sur les taux de reconnaissance.

On retrouve ici la constatation faite au paragraphe 632 : la classification globale guidée par le type des locutions-feuilles donne de meilleurs résultats en une propostion que la classification par type.

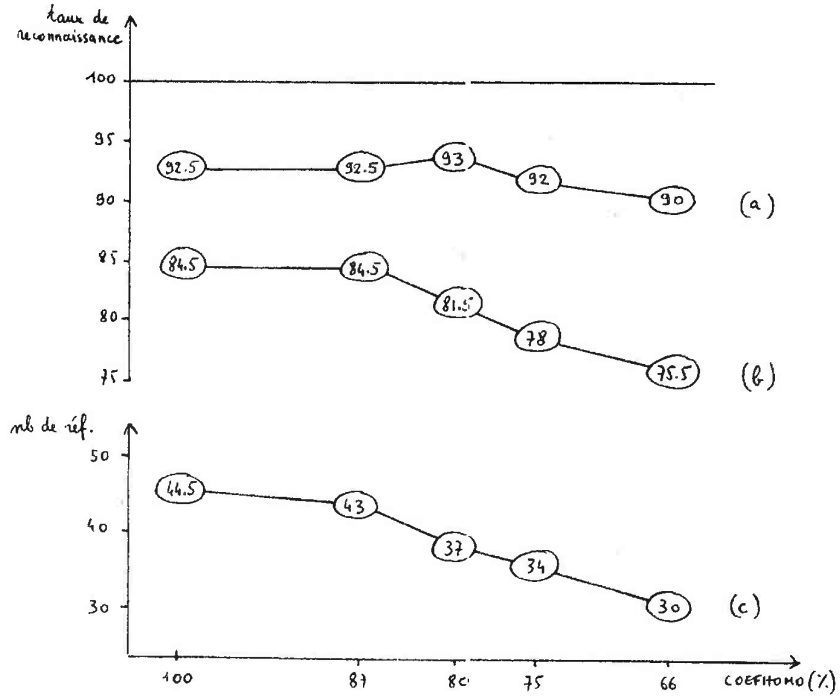
Parmi les différentes méthodes de classification par type, c'est la segmentation à niveau variable guidée par le nombre de points maximum par classe qui donne les meilleurs taux de reconnaissance .

Le nombre de classes, donc le nombre de références pour la reconnaissance, est une grandeur qu'il est utile de pouvoir contrôler puisqu'elle détermine le temps de reconnaissance . Seule la classification par type à niveau fixe permet un contrôle direct encore que peu souple. Pour les autres le contrôle se fait par l'intermédiaire d'un paramètre qu'il faut fixer à priori (nombre de points maximum par classe pour la classification par type, coefficient d'homogénéité pour la classification globale).

Dans les deux cas, réduire le nombre de références oblige à consentir une diminution du taux de reconnaissance . Par exemple pour la classification par type (cf. figure B42) pour économiser 13.5 références il faut accepter de perdre 3.5 points de reconnaissance (et 7.5 points en deux propositions). De même pour la classification globale (cf. figure B41) le taux décroit presque régulièrement de 79.5 à 74.3 au fur et à mesure que diminue le nombre de références (de 46 à 32).

Toutefois dans ce cas, un optimum parait se dégager vers 87% où le nombre de référence diminue de 46.2 à 44.5 sans influence sur le taux de reconnaissance.

Notons au passage qu'il ne s'agit pas seulement d'une moyenne, toutes les courbes obtenues ont très sensiblement la même allure. La décroissance peut être plus ou moins rapide mais l'optimum oscille toujours entre 87 et 80%. Par exemple la courbe obtenue avec la variance comme critère d'agrégation (courbe optimale) est donnée en figure B43.



(a) : taux de reconnaissance en deux propositions
(b) : taux de reconnaissance en une proposition
(c) : nombre de classes (pour 10 types de mots)

Figure B43: Classification globale, segmentation à niveau variable guidée par le type (critère de la variance).

64. CHOIX DU CENTROIDE ET DU RAYON D'INFLUENCE

641. But

Quels que soient les choix précédents (corpus, méthode (sauf CSA), segmentation, etc...) l'issue de la classification est une quasi-partition du corpus initial : c'est-à-dire que l'ensemble des locutions acquises lors de la phase d'apprentissage (à quelques exceptions près : les ''intrus'' éliminés lors de la segmentation) est réparti en un certain nombre de classes disjointes, chacune représentative d'une prononciation particulière d'un mot.

Le but de cette étape est de choisir pour chaque classe un représentant qui sera seul conservé pour la reconnaissance. Trois variantes différentes ont été envisagées :

- Centroide MINIMAX
- Centroide MINISOM
- Point moyen

642. Centroide MINIMAX

Définition : Le centroide MINIMAX d'un ensemble de points $E = \{x1, x2, \ldots xn\}$ est le point $X \in E$ qui vérifie :

$$\max_j (d(X, x_j)) = \min_i (\max_{j \neq i}(d(x_i, x_j)))$$

où $d(x_i, x_j)$ représente la pseudo-distance définie au paragraphe 62.

C'est le point qui présente la plus petite distance maximum aux autres points de l'ensemble. Le centroide minimax est un point proche du centre ''géométrique '' de l'ensemble, en effet cette définition ne tient compte que de l'enveloppe de l'ensemble et non de la répartition des points. (cf. figure B44).

643. Le centroide MINISOM

Définition : Le centroide MINISOM d'un ensemble de points $E = \{x1 , x2 ,\ldots.xn\}$ est le point $X \subset E$ qui vérifie :

$$\sum_j d(X,x_j) = \min_i(\sum_{j \neq i} d(x_i, x_j))$$

Figure B44: Centre Minisom (⊹) et Minimax (◎) d'un nuage de points.
On remarque que Minisom est plus proche de la zone dense

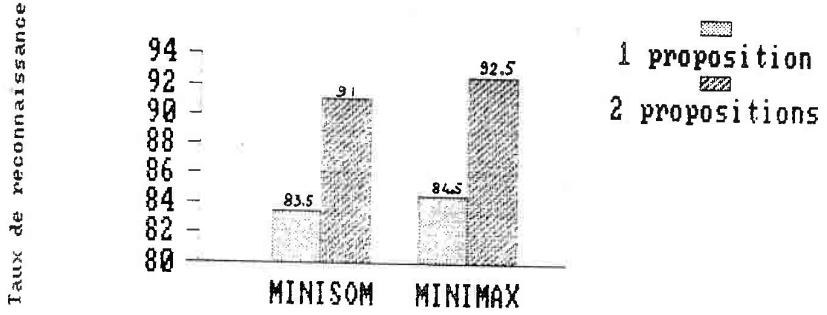


Figure B45: Comparaison des taux de reconnaissance
obtenus avec Minisom et Minimax

C'est le point qui présente la plus petite somme des distances aux autres points de l'ensemble. Le centroide Minisom est un point proche du centre de gravité de l'ensemble car sa définition fait intervenir la notion de répartition des points. (cf.figure B44).

644. Point moyen

Les deux définitions précédentes choisissent pour centroide, un point qui appartient à l'ensemble, la méthode du point moyen au contraire, fabrique un point par moyennage dynamique de tous les points de l'ensemble. On utilise pour cela un algorithme dérivé de celui décrit au paragraphe 62 [Briant 84]. Cette méthode n'est pas actuellement opérationnelle dans notre système, car elle ne semble pas a priori bien adaptée aux caractéristiques présentes du corpus. Celui-ci pouvant regrouper dans une même classe des élocutions sensiblement différentes d'un même mot, le représentant issu de ce moyennage ne présente aucune garantie de centralité géométrique. Des expériences de [Lockwood 85] semblent confirmer cette hypothèse: ''En moyennant des représentants d'un même mot, il n'est pas automatique que le mot résultant soit le minimax de l'ensemble. Ceci veut dire qu'il ne nous sera pas possible de faire systématiquement un moyennage pour avoir un prototype au centre de la classe ''.

645. Résultats

Les deux notions de centroide testées (Minisom et Minimax), donnent des résultats assez proches (cf. figure 345) mais privilégient toutefois légèrement le Minimax ce qui n'est pas surprenant compte-tenu du caractère purement ''topologique '' de la stratégie de reconnaissance utilisée (plus proche voisin et dérivés).

Le centre Minisom s'est toutefois révélé plus efficace que le centre Minimax lors de la recherche d'une région stable pour la méthode non hiérarchique (cf. 633 a). Son intérêt était en effet de diriger le déplacement des centres vers les zones de densité maximale ce qui répond bien à sa définition. (cf. B44). (la figure B25a a d'ailleurs été obtenue avec Minisom). Mais même dans ce cas une fois la classe stabilisée, le meilleur choix de représentant reste Minimax.

646. Rayon d'influence

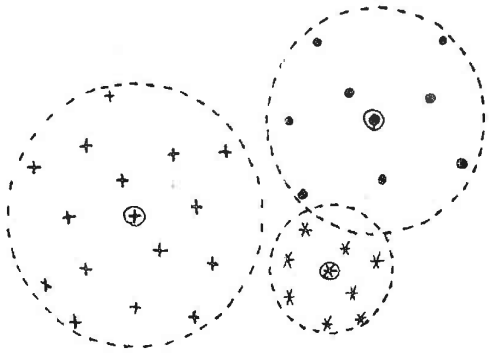
Le passage d'un ensemble de points (la classe de locutions) à un représentant (le centroide) entraîne une perte d'information (densité, volume, nombre de points de l'ensemble). Or, nous avons vu lors de la classification et au chapitre 5 que les classes présentaient de grandes différences de volume (types lâches ou denses), (c'est-à-dire d'indices d'agrégation pour la Classification Ascendante Hierarchique (cf. figure B46). Il nous a semblé intéressant de conserver cette information sous la forme du ''rayon d'influence '' d'un représentant défini comme suit :
Soit X0 le centroide représentant de l'ensemble E = {X1, X2, ..., Xn}

$$R = \max_{i} d(X0, Xi)$$

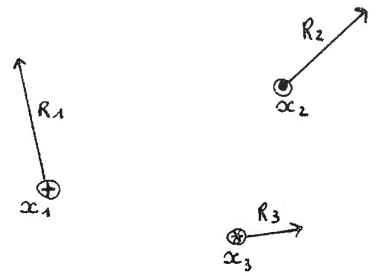
c'est la distance du centroide au point le plus éloigné de la classe qu'il représente.

Remarque 1 : Les références correspondant à des points isolés ont un rayon d'influence nul, ce qui les rendrait inutiles pour la reconnaissance ; c'est pourquoi, on leur affecte arbitrairement un rayon égal au plus petit rayon trouvé au cours de la classification.

Remarque 2 : Des tests de reconnaissance ont été faits avec et sans utilisation du rayon d'influence et prouvent la pertinence de la conservation de cette information. (cf. 665).



3 classes issues de l'étape de classification



Chaque classe représentée par son centroide et son rayon

Figure B46: Centroides et rayons sur un exemple fictif en 2 dimensions

65. TRAITEMENT DES SINGLETONS

651. Principe

Nous avons vu que la classification aboutissait généralement à un nombre de singletons (c'est-à-dire de classes réduites à une locution) assez important (cf. figure B40, paragraphe 635f).
Nous avons vu également qu'on pouvait en supprimer certains en jouant sur le taux de recouvrement des classes par le biais du coefficient d'homogénéité.

Lors de tests de reconnaissance effectués sur d'autres corpus nous avons constaté que l'efficacité de ces singletons était très variable, certains étant très utiles, c'est-à-dire ayant servi à plusieurs reconnaissance exactes, d'autres inutiles c'est-à-dire n'ayant jamais servi lors d'une reconnaissance, d'autres encore néfastes car n'ayant servi qu'à des reconnaissance inexactes (cf. figure B47).
Nous avons donc imaginé de faire subir à ces singletons un pré-test d'utilité sur le même corpus d'apprentissage qui a servi à la classification. Pour cela nous simulons une reconnaissance de chaque locution de l'ensemble d'apprentissage à l'aide des références issues de la classification, en mémorisant pour chaque référence le nombre de fois où elle s'est révélée :

- Utile, c'est-à-dire ayant servi à une reconnaissance exacte sans remplacement possible par une autre référence,
- Inutile, c'est-à-dire n'ayant pas servi à la reconnaissance ou ayant servi mais pouvant être remplacée par une autre,
- Néfaste, c'est-à-dire ayant servi à une mauvaise reconnaissance.

On conserve pour cela les trois meilleurs scores à chaque reconnaissance et on incrémente les notes d'utilité et de nuisibilité en fonction du tableau représenté figure B48 où :
= et ≠ signifient que le score correspond à une référence du même type (respectivement de type différent) que le mot prononcé.
Ui est la note d'utilité de la référence correspondant au i ème score;
Ni est la note de nuisibilitéde la référence correspondant au i ème score;

A l'issue de ce test chaque référence possède une note U et une note N, On réitère ensuite le processus après avoir supprimé la référence singleton ayant la plus forte différence N-U positive. On s'arrête lorsqu'on ne trouve plus de singletons répondant à ce critère.

STATISTIQUES PAR REFERENCE

BR1=Nb de fois ou cette référence a été bien reconnue en 1 ère position
MR1=Nb de fois ou cette référence a été mal reconnue en 1 ère position

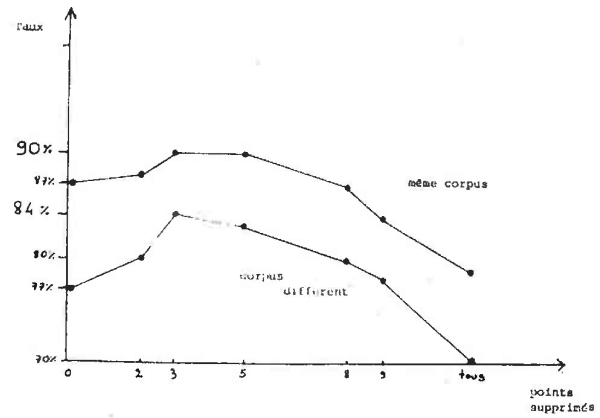
No Ref	No Locn	Type	Région	No Locn	BR1	BR2	MR1	MR2
1	110	DEUX	110	2	0	0	0	0
2	73	DEUX	102	7	2	0	0	0
7	14	NEUF	127	2	0	0	0	0
8	50	NEUF	110	5	4	2	0	0
13	63	SIX	105	6	1	0	1	0
14	39	SIX	127	4	2	0	0	0
19	28	QUATRE	120	3	2	0	0	0
20	16	QUATRE	117	2	1	0	0	0
24	29	UN	117	3	3	0	0	0
25	41	UN	117	4	0	0	2	1
28	30	HUIT	127	3	1	0	0	0
29	6	HUIT	110	1	5	2	2	1
33	55	SEPT	97	5	2	1	3	1
34	19	SEPT	112	2	3	0	0	0
38	44	CINQ	127	4	3	0	2	2
39	68	CINQ	127	6	1	0	0	0
43	45	TROIS	120	4	2	0	0	0
46	81	TROIS	117	7	3	1	0	0
47	94	ZERO	127	8	2	0	0	0
48	10	ZERO	112	1	2	0	0	0

La référence 7 est inutile (elle n'a jamais servi pour la reconnaissance)
La référence 8 est très utile (elle n'a servi qu'à des reconnaissances exactes)
La référence 25 est néfaste (elle n'a servi qu'à des reconnaissances fausses)
La référence 29 est plus utile que néfaste

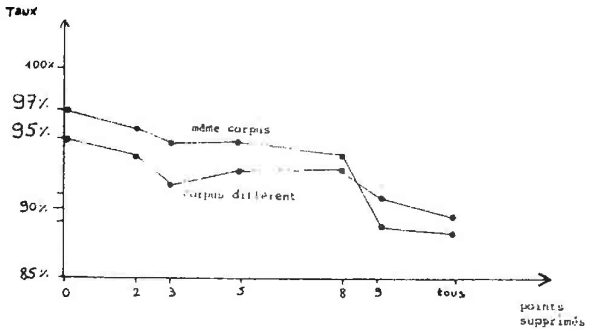
Figure B47: Exemples d'utilité des références

1er score	2eme score	3eme score	Action
=	=	=	-
=	=	≠	-
=	≠	=	U1=U1+1
=	≠	≠	U1=U1+2
≠	=	=	N1=N1+1
≠	=	≠	N1=N1+1 U2=U2+1
≠	≠	=	N1=N1+1 N2=N2+1
≠	≠	≠	-

Figure B48: Tableau de calcul des notes d'utilité



a) Evolution des taux en une proposition



b) Evolution des taux en deux propositions

Figure B49: Evolution des taux de reconnaissance en fonction du nombre de points supprimés

652. Résultats

Si l'on examine les résultats obtenus (cf. figure B49 a et b) quant aux taux de reconnaissance obtenus sur un corpus autre que celui d'apprentissage, on peut faire deux constatations :

* En supprimant les singletons par ordre décroissant de N-U, les taux de reconnaissance en première propositon augmentent dans un premier temps puis retombent assez rapidement ensuite.
* Les courbes corpus d'apprentissage et corpus différent évoluent parallélement, en particulier l'optimum du taux de reconnaissance a lieu pour le même nombre de singletons supprimés, ce qui justifie la méthode utilisée.

L'optimum ainsi fixé permet, tout en supprimant quelques références (ici 4) de gagner quelques points de reconnaissance (on passe de 81 à 84.5).

66. STRATEGIE DE RECONNAISSANCE

661. Introduction

A l'issue de toutes les étapes précédentes qui constituent l'apprentissage, nous avons donc obtenu un ensemble de références

P = {P1, P2, ..., Pn } (en moyenne quatre ou cinq par type de mot)

avec pour chaque référence Pi :

- Un numéro d'identification
- Son spectrogramme : Si
- Son type : Ti
- Son rayon d'influence : Ri

La phase de reconnaissance consiste à déclarer que le mot inconnu est du même type que la référence à qui il ressemble le plus.

Nous avons testé trois critères de ressemblance utilisant tous la notion de distance d (s_i, s_j) décrite au chapitre 62 :

* Critère du plus proche voisin (PPV)
* Critère du rayon (RAY)
* Critère de ''l'intériorité'' (INT)

Soit l'ensemble de référence :

P = {(S1, T1, R1), (S2, T2, R2), ..., (Sn ,Tn ,Rn)}

et le mot à reconnaître :

P0 = (S0 ,T0)

où S0 est le spectrogramme acquis

T0 son type (à déterminer)

662. Critère du plus proche voisin

Définition :

T0 = T* ∈ (S*, T*, R*) tel que

d(S0 ,S*) = Min (d(S0 ,Si))

Le mot inconnu est assimilé à la référence la plus proche, c'est-à-dire celle qui réalisera le meilleur score de comparaison dynamique. Pour cette méthode on ne tient pas compte du rayon. (cf. figure B50a).

663. Critère du rayon

Définition :

T0 = T* ∈ (S* ,T* ,R*) tel que

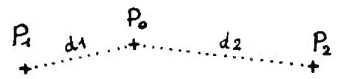


Figure B50a: Critère du plus proche voisin. (Le mot inconnu P0 est assimilé à la référence P1 car d1<d2)

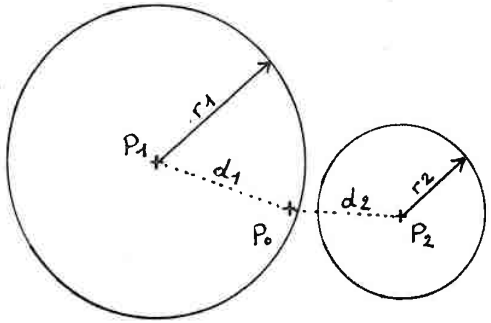


Figure B50b: Critère du rayon. (Le mot inconnu P0 est assimilé à P1 bien que plus proche de P2 car d1<r1 et d2>r2)

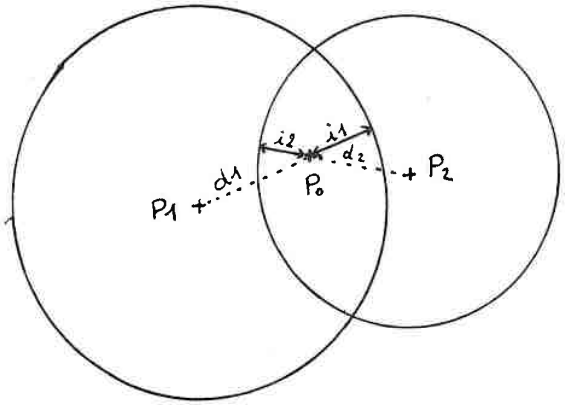


Figure B50c: Critère de l'intériorité. (Le mot inconnu P0 est assimilé à P1 car, bien que plus proche du centre P2, il est plus loin de la périphérie de P1 : i1>i2)

d(S0 ,S*) = Min (d(S0 ,Si)) vérifiant
d(S0 ,S*) ≤ Ri

Le mot inconnu est assimilé à la référence la plus proche à condition qu'il se situe à l'intérieur de la sphère d'influence de celle-ci (c'est-à-dire à une distance inférieure à son rayon).(cf. Figure B50b).

664. Critère de l'intériorité

T0 = T* c (S* ,T* ,R*) tel que
d(S0 ,S*) = Min (d(S0 ,Si) - Ri)

Cette notion privilégie au caractère ''proche du centre '', le caractère qui présente la limite d'influence la plus éloignée. (cf.figure B50c).

665. Résultats

On voit, sur la figure B51, que les résultats les meilleurs sont obtenus par les deux critères qui prennent en compte le rayon ce qui confirme l'hypothèse du paragraphe 646. Parmi eux c'est l'intériorité qui fournit les meilleurs taux de reconnaissance . Ceci peut s'expliquer par une ligne de démarcation plus pertinente dans le cas de classes sécantes.(cf. Figure B52).

Remarque : Il s'est révélé interessant de disposer d'un coefficient d'''expansion'' des sphères afin de pouvoir adapter au mieux l'ensemble de références munies de leur rayon au corpus de reconnaissance . Ce paramètre agit en multipliant tous les rayons d'influence des références par un coefficient constant. Le coefficient optimum s'est curieusement révélé très constant au cours des tests. Il est de 1.1 pour le critère du rayon et de 1.4 pour l'intériorité.(cf.figure B53).

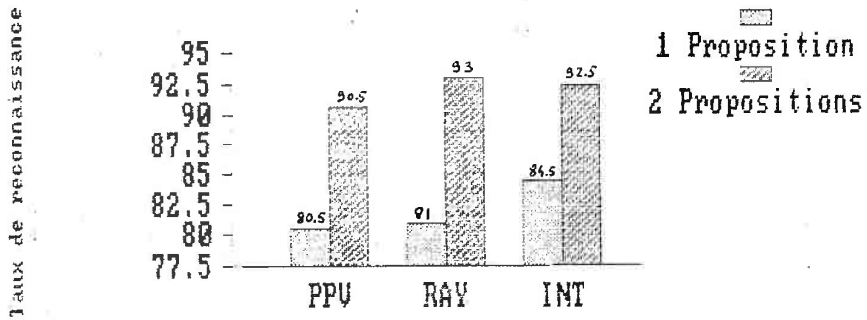


Figure B51: Comparaison des taux obtenus avec les 3 critères de reconnaissance

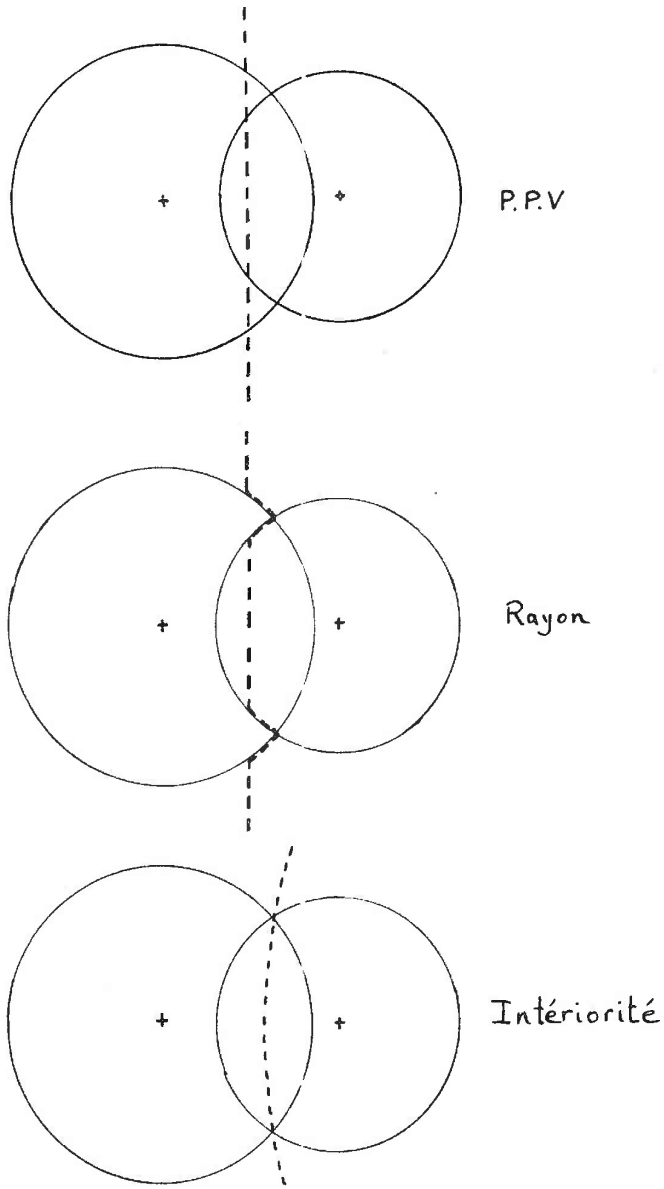
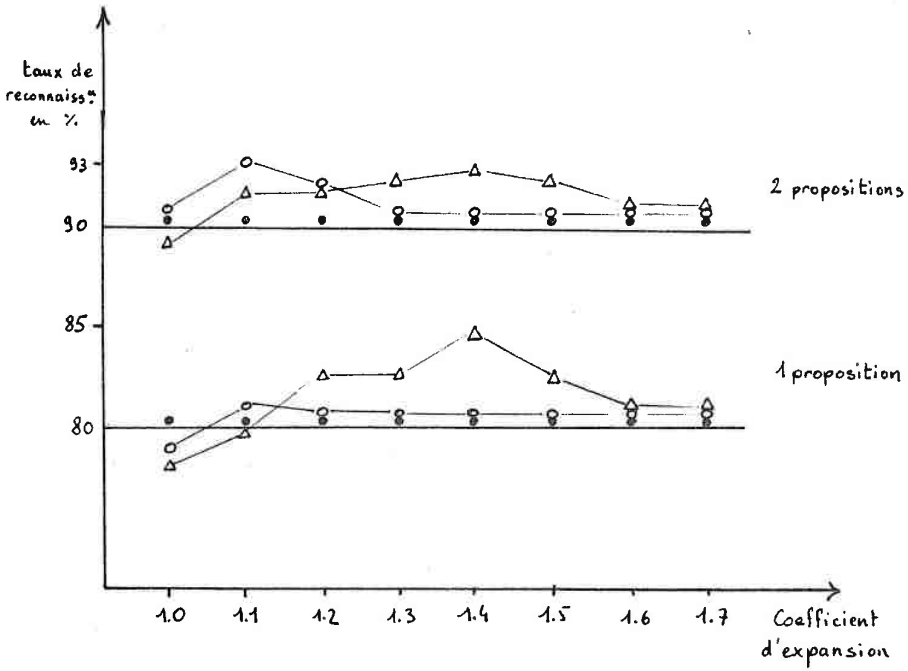


Figure B52: Lignes de démarcation dans le cas de 2 sphères sécantes, pour les trois critères de reconnaissance



- Critère du P.P.V
- Critère du rayon
- Critère de l'intériorité

Figure B53: Evolution des taux de reconnaissance en fonction du coefficient d'expansion des sphères

67. LA METHODE C.S.A

671. Un arbre de référence

Nous détaillerons ici la méthode de représentation C.S.A. (Conservant la Structure d'Arbre), traitée à part car elle ne suit pas toutes les étapes présentées lors de la structure de base (cf, Chapitre 4). En particulier, bien que basée sur une C.A.H., elle ne segmente pas l'arbre hiérarchique donc les notions de classes, de points isolés, de centroide n'y ont pas de (ou pas la même) signification et les stratégies de reconnaissance y sont très différentes.

a) Description

C.S.A n'est pas une méthode de réduction des données à proprement parler mais une méthode de structuration des données (cf partie A paragraphe 444). En efet c'est le corpus d'apprentissage entier qui va servir lors de la reconnaissance mais structuré de telle sorte que la comparaison du mot inconnu à cette structure de référence soit plus rapide qu'une comparaison séquentielle de toutes les locutions qui y figurent.

La structure choisie est celle d'un arbre hiérarchique obtenu de la façon décrite au paragraphe 633b mais dans lequel chaque noeud est étiqueté par trois valeurs :

- un représentant qui est le centroide de toutes feuilles issues du noeud.
- Le rayon d'influence de ce représentant calculé comme au paragraphe 64.
- Son type si toutes les feuilles issues du noeud sont de même type.

où les deux fils du noeud (aîné et benjamin) si toutes les feuilles issues de sont pas de même type. Il s'agit bien sûr ici d'une classification globale (c'est-à-dire tous types confondus).

b) Algorithme d'étiquetage

Cet algorithme récursif utilise l'arbre issu de la C.A.H et ajoute une étiquette à chacun de ses noeuds, le résultat est un tableau (TARB) équivalent à l'arbre étiqueté.

Il est initié par l'instruction Etiquetage(TR) où TR est le tronc de l'arbre.

Etiquetage (Noeud)

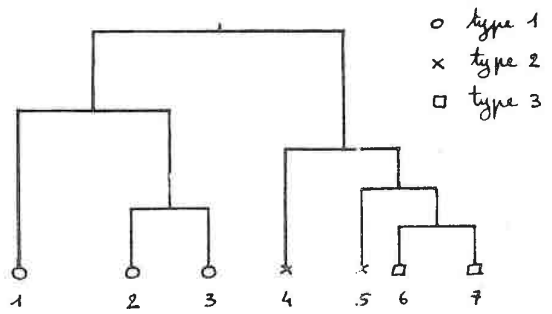
- Lister toutes les feuilles issues de (Noeud)
- Calculer le centroide de ces feuilles et le rayon d'influence.
- Si toutes les feuilles sont du même type

alors associer à Noeud dans le tableau TARB
- son représentant (le centroide)
- son rayon d'influence
- son type

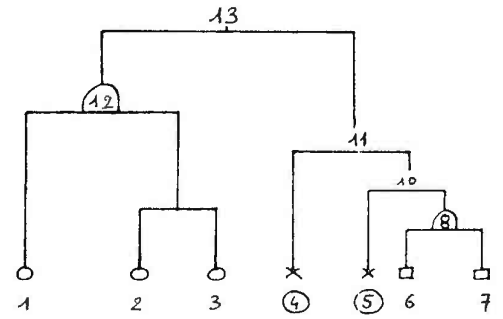
sinon associer à Noeud dans le tableau TARB
- son représentant (le centroide)
- son rayon d'influence
- son fils ainé
- son benjamin

Etiquetage (Ainé (NOEUD))
Etiquetage (Benj (NOEUD))

Prenons comme exemple l'arbre hiérarchique suivant:



Il sera étiqueté de la façon suivante:



où seuls les noeuds entourés sont des noeuds typés.

Le tableau TARB équivalent sera:

n° point	Représentant	Rayon	Type	Ainé	Benj.
13	SP_{13}	218	-1	12	11
12	SP_{12}	121	1	-1	-1
11	SP_{11}	102	-1	10	4
10	SP_{10}	92	-1	8	5
8	SP_8	70	3	-1	-1
5	5	0	2	-1	-1
4	4	0	2	-1	-1

où les SPi sont les spectrogrammes des centroides-représentants (lorsqu'un point est une feuille, celle-ci est son propre représentant).

c) Reconnaissance

Le mot inconnu au lieu d'être comparé séquentiellement à toutes les références comme dans les méthodes précédentes va ici ''descendre '' les branches de l'arbre de référence de façon dichotomique depuis le tronc jusqu'au premier noeud typé.
A chaque noeud, deux branches se présentent, l'algorithme suivra celle dont le représentant est le plus proche du mot à reconnaitre.
Les notions de proximité sont toujours issues des algorithmes de comparaison dynamique présentés au paragraphe 62.
Puisque seul le type importe, lors de la reconnaissance une fois arrivé à un noeud typé, il n'est pas utile de descendre l'arborescence plus loin, c'est pourquoi dans le tableau TARB les noeuds typés sont considérés comme n'ayant pas de descendants (cf. noeud 12 de l'exemple ci-dessus).

Algorithme de reconnaissance :
Algorithme récursif initié par l'instruction:
COMPARB (Spectro, Tr, Refrec, Score) où
Tr est le tronc de l'arbre.
Spectro est le spectrogramme du mot inconnu,
Refrec est la référence reconnue,
Score est le score de reconnaissance final,
Représentant est le spectrogramme du représentant-centroide.
COMPARE est la procédure de comparaison dynamique de deux spectrogrammes.

```
COMPARB (Spectro, Noeud, Refrec, Score)

  - COMPARE (Spectro, Représentant(Ainé(Noeud))) ---> Scor1
  - COMPARE (Spectro, Représentant (Benj(Noeud))) ---> Scor2

  - Si Scor1 < Scor2  alors | Fils = Ainé(Noeud)
                            | Scoref = Scor1
                      sinon | Fils = Benj (Noeud)
                            | Scoref = Scor2

  - Si type (Fils) ≠ -1   alors | Refrec = Fils
                                | Score = Scoref
                          sinon  COMPARB (Spectro, Fils, Refrec, Score)
```

d) Résultats

Cet algorithme a été testé et donne de mauvais résultats (<60%) ; l'analyse des erreurs montre en effet que l'aiguillage dans l'arborescence lors des quelques premiers noeuds n'est absolument pas fiable. Ceci est dû au fait que la notion de représentant n'est plus significative pour un grand nombre de locutions très différentes ; en effet les premiers représentants, au sommet de l'arbre, sont des centroides d'un ensemble d'au moins 50 locutions réparties en au moins cinq types différents.

Les premiers représentants étant non significatifs, les premiers choix d'exploration de l'arbre sont donc pratiquement aléatoires ce qui explique les mauvais taux de reconnaissance . Par contre nous avons pu remarquer que dans le cas d'un choix correct lors des deux premiers aiguillages, le reste de la reconnaissance était bien conduit et aboutissait beaucoup plus souvent à une réponse exacte.

C'est pourquoi nous avons choisi d'adapter cet algorithme de deux façons différentes, mais toutes deux guidées par la nécessité de ne pas contraindre à des choix définitifs trop tôt, tout en veillant à ce que la notion de représentant reste pertinente.

672. Une forêt de référence

Les deux algorithmes suivants constituent un moyen terme entre deux méthodes opposées : dichotomique complète et séquentielle complète. Les locutions de référence sont cette fois structurées en une forêt (c'est-à-dire un ensemble d'arbres). La reconnaissance aura ainsi la forme d'une séquence d'explorations dichotomiques où chaque arbre sera suffisamment réduit et cohérent pour que la notion de représentant y reste fonctionnelle.

a) Classification par type

* Apprentissage

Il s'agit ici de construire un arbre par type de mot afin d'assurer la cohérence des feuilles-locutions. Nous allons donc appliquer, à chaque ensemble de locutions de même type un algorithme d'étiquetage semblable à celui vu précedemment mais sans considération de type, l'étiquetage est donc poursuivi jusqu'aux feuilles.


```
ETIQUETAGE-T (Noeud)

Si NOEUD est une feuille

    alors associer à Noeud dans le tableau TARB
            *son spectrogramme
            *son type

    sinon
        - lister les feuilles issues de Noeud
        - calculer leur centroide et leur rayon
        - associer à Noeud dans le tableau TARB
            * son représentant
            * son rayon
            * son fils ainé
            * son fils benjamin
        - ETIQUETAGE-T(Ainé(Noeud))
        - ETIQUETAGE-T(Benj(Noeud))
```

A l'issue de l'étape d'apprentissage, nous obtenons donc un arbre par type.

* Reconnaissance :

Lors de la reconnaissance, le mot inconnu est comparé séquentiellement à chacun des arbres avec un algorithme proche de COMPARB mais descendant jusqu'aux feuilles sans considération de type. On obtient ainsi une référence par type, le mot à reconnaitre est assimilé à celle qui a réalisé le meilleur score.

Algorithme:

```
RECONNAISSANCE-T

    Scormin = ∞
    Pour i de 1 à Nb_de_types    répéter
        | COMPARB-T (Spectro, Tr_i , Refrec_i , Score_i )
        | Si Score < Scormin  alors
                        | Référence_reconnue = Refrec_i
                        | Scormin = Score_i
```

où COMPARB-T est équivalent à COMPARB en remplaçant la condition

- Si type (Fils) ≠ -1
  par
- Si FILS est une feuille

b) Classification globale

L'ensemble d'arbres est ici obtenu à partir de l'arbre hiérarchique complet de toutes les locutions, tous types confondus, celui-ci étant segmenté horizontalement de façon à isoler le nombre voulu de branches. (Méthode analogue à "segmentation à niveau fixe guidée par le nombre total de classes" (cf. paragraphe 635). Dans ce cas les branches isolées (devenues donc des arbres) peuvent regrouper plusieurs types différents, mais elles sont supposées relativement homogènes du point de vue des distances entre locutions, ce qui permet d'y utiliser la notion de représentant.

Chaque arbre est ensuite étiqueté à l'aide de l'algorithme "Etiquetage-t". Le nombre d'arbres est un paramètre du système : plus il est grand, plus la méthode s'apparente à une recherche séquentielle (avec augmentation du temps de reconnaissance) mais plus la notion de représentant y est fonctionnelle.

Lors de la reconnaissance le mot inconnu est comparé séquentiellement à la "forêt de référence " avec l'algorithme Comparb-t. On obtient ainsi autant de références que d'arbres, avec éventuellement plusieurs références du même type et certains types non représentés. Le mot à reconnaitre est assimilé à la référence qui a réalisé le meilleur score à l'aide d'un algorithme équivalent à Reconnaissance - t.

c) Résultats avec "forêt de référence "

On constate sur la figure B54 que les taux de reconnaissance ont été nettement améliorés ce qui confirme l'analyse des erreurs faite au paragraphe 671. Ils restent cependant moins bons qu'avec les méthodes séquentielles, principalement en première proposition. On remarque également qu'ici encore, la méthode globale donne de meilleurs résultats que la méthode par type.

Nous avons vu que pour les méthodes séquentielles, il était intéressant de faire intervenir la taille des classes par le biais d'un rayon d'influence et que cela influait favorablement sur les taux de reconnaissance. Dans tous les algorithmes C.S.A présentés cette possibilité a été gardée, mais lors des tests, c'est le P.P.V. qui a été utilisé, il serait intéressant de voir comment la notion de rayon pourrait faire évoluer les résultats de reconnaissance.

De plus, l'avantage d'une méthode dichotomique serait peut être plus net sur un corpus plus étendu.

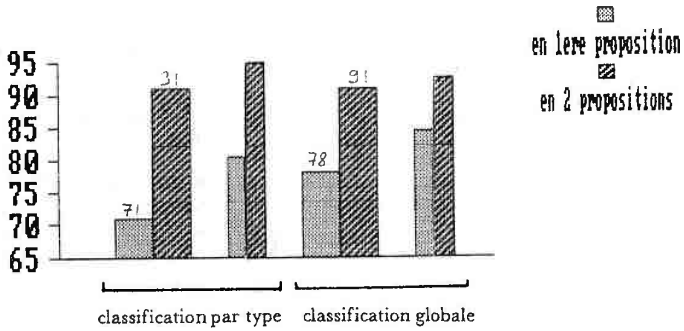


Figure B54: Taux de reconnaissance utilisant une ''forêt de référence '' composée de 10 arbres.
Les petites colonnes représentent pour mémoire les taux obtenus avec l'optimum des méthodes séquentielles.

d) Comparaison des temps de reconnaissance

N est le nombre de locutions du corpus d'apprentissage
T est le nombre de types
C est le nombre de références voulues par type
D est le nombre de comparaisons à faire par mot à reconnaitre

Théorique	dans le cadre de l'application (N=100)
* Séquentiel	
Dmin = Dmax = Dmoy = C.T	40 à 50
* Dichotomique	
- Arbre parfaitement dichotomique	
2. log N	13 ou 14
- Arbre ''en chaine ''	
Dmin = 2	Dmin = 2
Dmax = 2(N-1)	Dmax = 198
Dmoy = $2.(\frac{N(N-1)}{2} - 1)/N$	Dmoy ≃ 100
≃ N	
- Arbre réel moyen	Dmoy = 14.9
* Mixte (K est le nombre d'arbres)	
- Arbre parfaitement dichotomique	10 arbres = 66.4
Dmoy = 2.K.log	5 arbres = 43.2
	4 arbres = 37.1
	3 arbres = 30.3
- Arbre réel moyen	10 arbres = 68
	5 arbres = 47.3
	4 arbres = 38
	3 arbres = 30

e) Conclusion

Le tableau précédent amène deux constatations:
Le nombre de comparaisons effectuées réellement est très proche des nombres théoriques calculés pour des arbres parfaitement dichotomiques, ce qui prouve que les arbres obtenus lors de la CAH sont ''bien formés'' et donc assez bien adaptés à ce type de méthode.

La reconnaissance mixte (c'est à dire avec forêt de références) étant la seule dont les taux soient acceptables, on constate que la méthode CSA est inadaptée à la taille de notre application; en effet, avec 10 arbres, elle provoque plus de comparaisons (68) que la méthode séquentielle (43) avec de moins bons résultats. On peut d'ailleurs constater en toute généralité que la méthode dichotomique mixte ne devient intéressante (en nombre de comparaisons) par rapport à la methode séquentielle que lorsque

$$\log_2\left(\frac{N}{T}\right) < \frac{C}{2}$$

si on impose au moins un arbre par type.

Par exemple, pour 1000 locutions réparties en 10 types, la méthode CSA aboutit en moyenne au même nombre de comparaisons qu'une méthode séquentielle avec 13 références par type mais avec la possibilité d'accès à la totalité du corpus d'apprentissage.

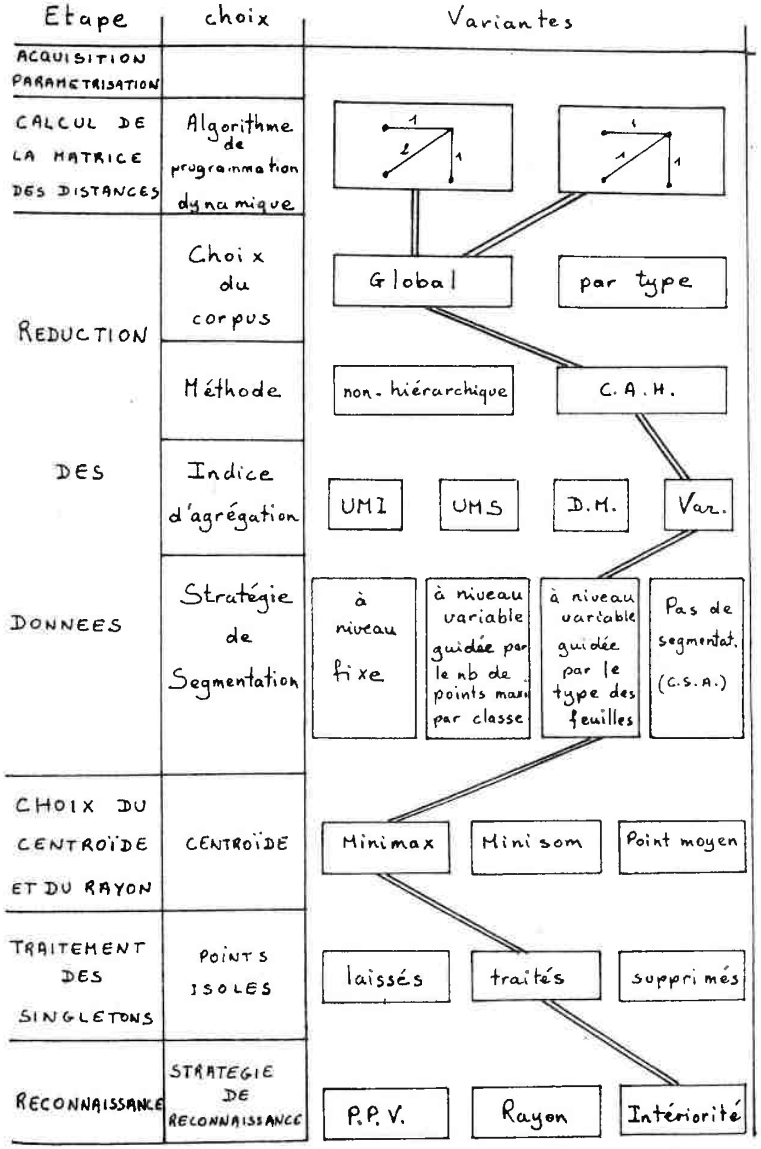
CHAPITRE 7

ANALYSE DES RESULTATS

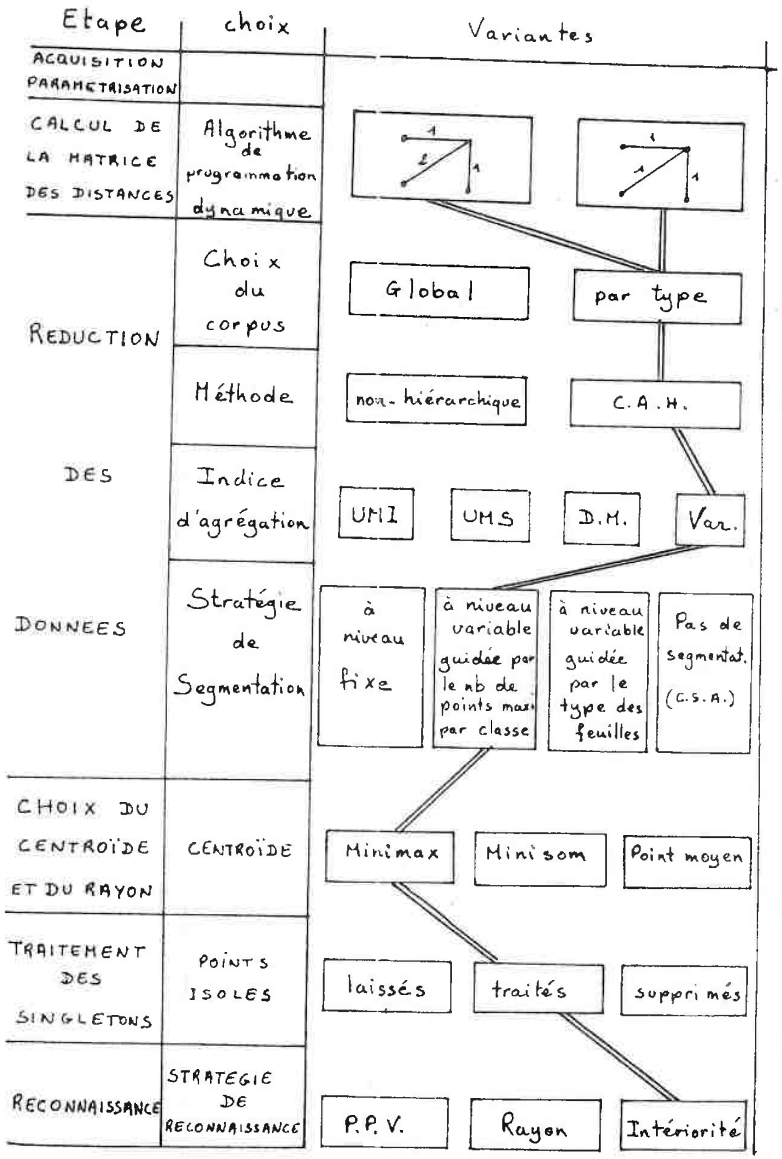
71. RESULTATS DE RECONNAISSANCE

711. Corpus C20

Les deux synoptiques suivants récapitulent les différentes variantes pour chaque mode de classification (par type et globale) et leur chemin optimal ainsi que les taux de reconnaissance obtenus en une et deux propositions sur un corpus C20 c'est-à-dire 10 locuteurs pour l'apprentissage et 10 locuteurs différents pour la reconnaissance.



Chemin optimal avec la classification globale : (84.5 % ; 92.5 %)



Chemin optimal avec la classification par type : (80.5 % ; 95 %)

Un certain nombre de constatations assez stables au cours de ces tests peuvent être faites:

- La classification globale avec segmentation guidee par le type des locutions donne presque toujours de meilleurs résultats que la classification par type pour les raisons citées au paragraphe 632.

- Parmi les quatre critères d'agrégation testés, la variance est celle qui donne les meilleurs résultats et les plus fiables car les plus constants: Au cours des tests effectués c'est le critère qui a présenté le plus faible écart (3%) contre 5.7% pour UMS et 5.3% pour DM.

- De plus, nous avons mis en évidence que la conservation et la prise en compte lors de la reconnaissance, d'une information sur la taille des classes permettait, dans un corpus assez hétérogène, une amélioration sensible des taux de reconnaissance, l'optimum étant atteint avec le critère de ''l'intériorité'' qui présente un gain de 4 % par rapport au PPV.

D'un point de vue plus général, le système MIMULE obtient des performances acceptables étant donnés le contexte multilocuteur et la grande hétérogénéité du corpus:

84.5% en une proposition

93% en deux propositions; avec 4.3 références par type de mot en moyenne.

Afin de situer ces performances relativement à la quantité d'informations retenues nous avons effectué deux types de tests de reconnaissance:

- L'un avec 40 références prises au hasard; on obtient: (66%; 74.5%).
- L'autre avec la totalité du corpus d'apprentissage (100 références), servant pour la reconnaissance; on obtient: (86% , 98%).

On constate la pertinence de la CAH qui avec 43% du nombre de références conserve 98% de l'information utile.

Afin d'éprouver la stabilité de ces conclusions, nous avons effectué certains tests complémentaires sur deux corpus plus grands l'acquisition-paramétrisation ayant été réalisée sur Masscomp à l'aide de la chaine décrite au paragraphe 432.

712. Corpus C44

1760 locutions réparties en:

44 locuteurs,

10 types de mots,

4 répétitions de chaque type,

L'apprentissage est fait sur 4 répétitions de 24 locuteurs (12 hommes et 12 femmes) pour la classification par type; et pour la classification globale, sur 2 répétitions de 16 locuteurs (8 hommes et 8 femmes) par type (soit 51040 comparaisons prenant environ 4 heures). La reconnaissance porte sur 20 autres locuteurs (14 hommes et 6 femmes).

Les résultats principaux sont les suivants:

Classification par type (Variance, PPV)

- Segmentation à niveau fixe guidée par le nombre de classes: 68.4% / 85% (50 références)
- Segmentation à niveau variable guidée par le nombre maximum de points par classe: 83.4% / 89.6% (46 références)

Classification globale: Variance, coefficient de recouvrement=0.8, élimination de tous les points isolés (Card Ci < 3), intériorité. (33 références):

- Tests multilocuteurs: 79.8% / 88.7%
- Tests plurilocuteurs: 87.1% / 96.7%

Ici encore, la variance donne des résultats supérieurs aux autres critères; en moyenne 2.1% de mieux que UMS et 3.4% de mieux que DM.

La différence essentielle révélée par ce corpus par rapport au corpus C20, porte sur le critére de reconnaissance. Les critères utilisant le rayon n'améliorent pas, voire dégradent légèrement la reconnaissance dans le cas d'une classification par type (-0.8% en moyenne par rapport au PPV). Dans le cas de la classification globale, au contraire, l'intériorité reste le critère optimal. (79.8% / 88.7% au lieu de 77.7% / 85.8%).

713. Corpus C60

Ce corpus utilise 2400 locutions réparties en:
- 60 locuteurs
- 10 types
- 4 répétitions par type

Pour les tests multilocuteurs l'apprentissage est réalisé avec les 4 répétitions des 30 premiers locuteurs (15 hommes, 15 femmes), la reconnaissance avec les 30 autres (15 hommes, 15 femmes également).

Pour les tests plurilocuteurs l'apprentissage est fait avec 2 répétitions de chacun des 60 locuteurs, la reconnaissance utilise les 2 autres.

Les résultats principaux sont les suivants:

classification par type (Var, PPV)
- Segmentation à niveau fixe guidé par le nombre de classes:
(52 références) 80.6 / 88.7
- Segmentation à niveau variable guidée par le nombre de points maximum par classe:
(74 références) 79.1 / 86.5

Classification globale (Var, coefficient de recouvrement=0.8, intériorité):
(50 références) 73.3 / 85.0

Notons qu'une véritable classification globale est ici impossible à cause du volume des informations mises en jeu, nous avons donc eu recours à un processus en deux temps: Une classification par type sélectionne 20 classes par type, parmi celles-ci on conserve les 15 plus grosses dont on extrait un représentant. On effectue ensuite une classification globale sur les 150 représentants tous types confondus.

72. ANALYSE DES TAUX OBTENUS

On constate que les taux de reconnaissance évoluent en fonction inverse de la taille du corpus. Ceci est un phénomène connu dû au fait qu'un grand corpus de reconnaissance présente une plus grande variété dans les prononciations; ce phénomène, selon Rabiner, s'atténue lorsque le corpus d'apprentissage dépasse 50 locuteurs.

De plus, les taux peuvent paraître assez faibles en moyenne par rapport aux résultats anoncés pour d'autres systèmes multilocuteurs (cf Partie A). Outre le fait que les taux sont très difficilement comparables lorsqu'ils sont testés dans des conditions différentes, rappelons que notre but n'était pas de faire un système compétitif mais de comparer entre elles diverses méthodes de classification et leur variantes. On peut cependant, après analyse des résultats, avancer les raisons suivantes:

721. Difficulté du vocabulaire

Le vocabulaire des chiffres français est un vocabulaire difficile (monosyllabique, proximité de certains chiffres comme 5 et 7) et il n'est complété d'aucun 'mot de service ' généralement mieux reconnu.

De plus, la segmentation du signal en mots est effectuée automatiquement ce qui entraîne des erreurs de localisation des débuts et fins de mots dans les cas de fricatives et plosives (cf chapitre 5); aucune correction ni élimination manuelle n'étant effectuée, certains chiffres présentent de fréquentes confusions (voir la matrice de confusion figure B55).

Afin de confirmer cette explication, nous avons effectué des tests de reconnaissance sur deux sous-vocabulaires séparés: Un vocabulaire difficile (les chiffres de 5 à 9) et un vocabulaire plus simple car mieux différencié (les chiffres de 0 à 4); le premier obtient 77/97% contre 93/100% pour le second. Une telle différence confirme l'influence du choix du vocabulaire sur les taux de reconnaissance. Rappelons également qu'on obtient 100% de reconnaissance sur le vocabulaire 'oui/non' qui est bien différencié.

722. Corpus de locuteurs

Précisons d'abord que les tests ont été effectués en ''indépendant du locuteur'' c'est-à-dire qu'aucun locuteur de la reconnaissance n'a participé à l'apprentissage. On constate sur la figure B56 une amélioration nette des taux de reconnaissance (+ 5 points en moyenne) lorsqu'on fait la reconnaissance sur les mêmes locuteurs qu'à l'apprentissage (tests

plurilocuteurs). En outre, les locuteurs n'ont pas été choisis et n'ont pas pour la plupart l'habitude du micro, ce qui introduit quelques effets indésirables lors de l'acquisition (claquement de langue, souffle, distance au micro...).

D'autre part, le corpus regroupe des accents bien différents, dont certains assez prononcés (lorrain, mosellan, maghrebin). Cette grande diversité des locutions influe sur l'homogénéité de la classification et peut perturber la reconnaissance. On constate par exemple sur la figure B57 que trois locuteurs ont réalisé un score inférieur à 70% en première proposition; il y a parmi eux une locutrice d'origine algérienne et un locuteur d'origine scandinave. On constate également (figures B57, B58) que 90% de la population obtient en moyenne 92% de reconnaissance exacte en deux propositions, à l'inverse on notera qu'on n'obtient pas de différence significative entre hommes et femmes (moins de 1%).

723. Algorithme de comparaison

Pour les corpus C44 et C60, un certain nombre de paramètres étaient imposés indépendemment de notre système : Principalement les algorithmes de segmentation du signal avec séparation signal-bruit, l'étape de paramétrisation (FFT) et l'algorithme de comparaison dynamique. Ce dernier, en particulier est assez rudimentaire (pas de contrainte de pente ni de normalisation de l'énergie). Il semble que cette chaine d'acquisition se révèle peu capable de gérer correctement la notion de distance entre les mots dans un contexte multilocuteur. Il est en effet significatif que la ''force brute '' c'est-à-dire la reconnaissance faite avec la totalité du corpus d'apprentissage sans classification ne donne pas un taux supérieur à 76%/89%. De même des tests de reconnaissance inter-locuteurs donnent de mauvais résultats: La reconnaissance du vocabulaire d'un locuteur à l'aide du vocabulaire d'un autre locuteur de même sexe pris au hasard aboutit à un score moyen de 52% en première proposition. Il est manifeste que la notion de distance entre mots utilisée respecte mal la notion humaine. On constate par exemple que ''un'' et ''neuf'' qui à l'oreille n'ont rien de commun sont fréquemment mesurés comme proches par l'algorithme de comparaison dynamique (plus proches quelquefois que deux ''neuf'' entre eux).

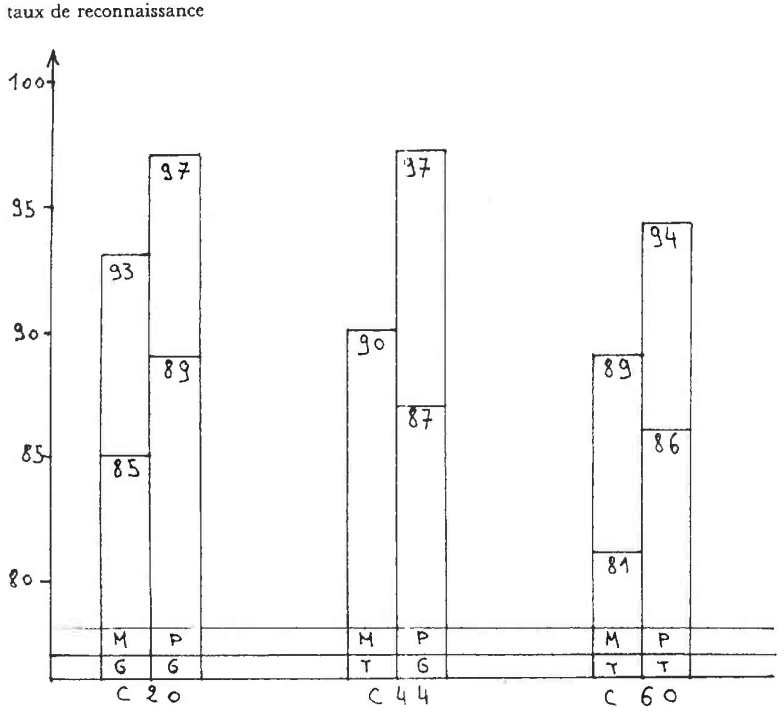
Si l'on se fixe pour but d'améliorer les scores il semble que cela soit possible en utilisant, pour la comparaison, des notions de base plus ''physiologiques'': cepstre mel, modèle d'oreille, ''perceptive linear prediction'' [Wakita 86].

D'autre part, le contexte de l'application nous a amené à imposer une décision de reconnaissance dans tous les cas, puisque de toutes façons

on demande confirmation du chiffre reconnu. Si l'on se fixe un seuil de confiance en dessous duquel on ne prend pas de décision, les scores sont légèrement améliorés: Avec 7% de rejet on obtient un score de 83% en une proposition et 91% en deux propositions. Mais même dans ce cas, il est significatif de remarquer que les chiffres les plus mal reconnus (5 et 7) ne bénéficient pas de cette amélioration, les confusions réalisant en général un bon score.

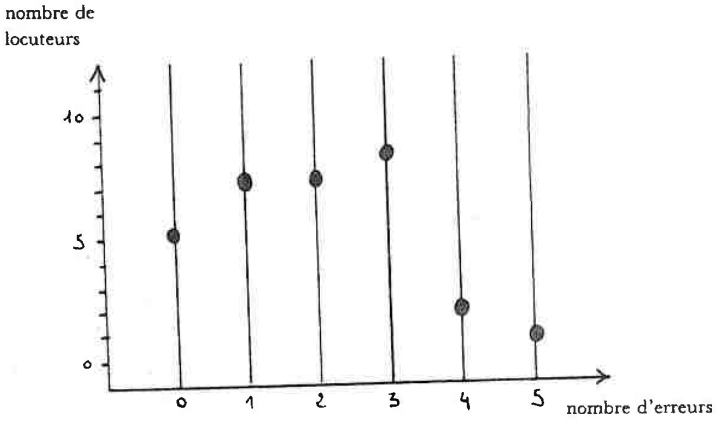
Mot prononcé \ Mot reconnu	0	1	2	3	4	5	6	7	8	9
0	97		3							
1		77	7	7						9
2	3		88	3		3				3
3				93		4				3
4		3	4	3	73				4	13
5	10	3			10	60		17		
6	7		3				73	4	13	
7	10		3		10	14		53		10
8									100	
9			4		3					93

Figure B55: Matrice de confusion sur le corpus C60 en première proposition (pourcentages)

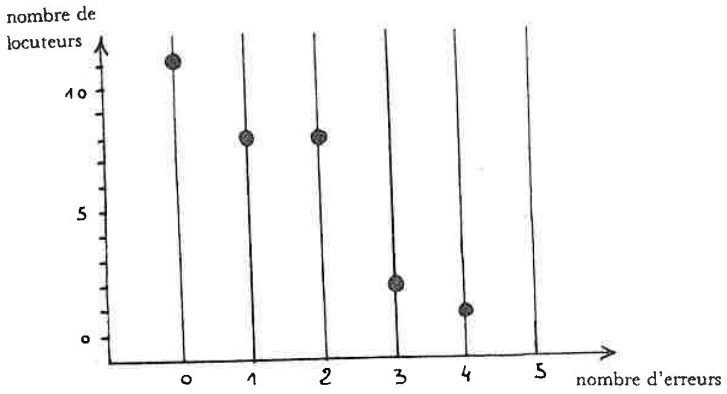


M : multilocuteur
P : plurilocuteur
G : classification globale
T : classification par type

Figure B56: Comparaison des reconnaissances multilocuteur et plurilocuteur



En 1 proposition



En 2 propositions

Figure B57: Nombre d'erreurde reconnaissance par locuteur en 1 et 2 propositions

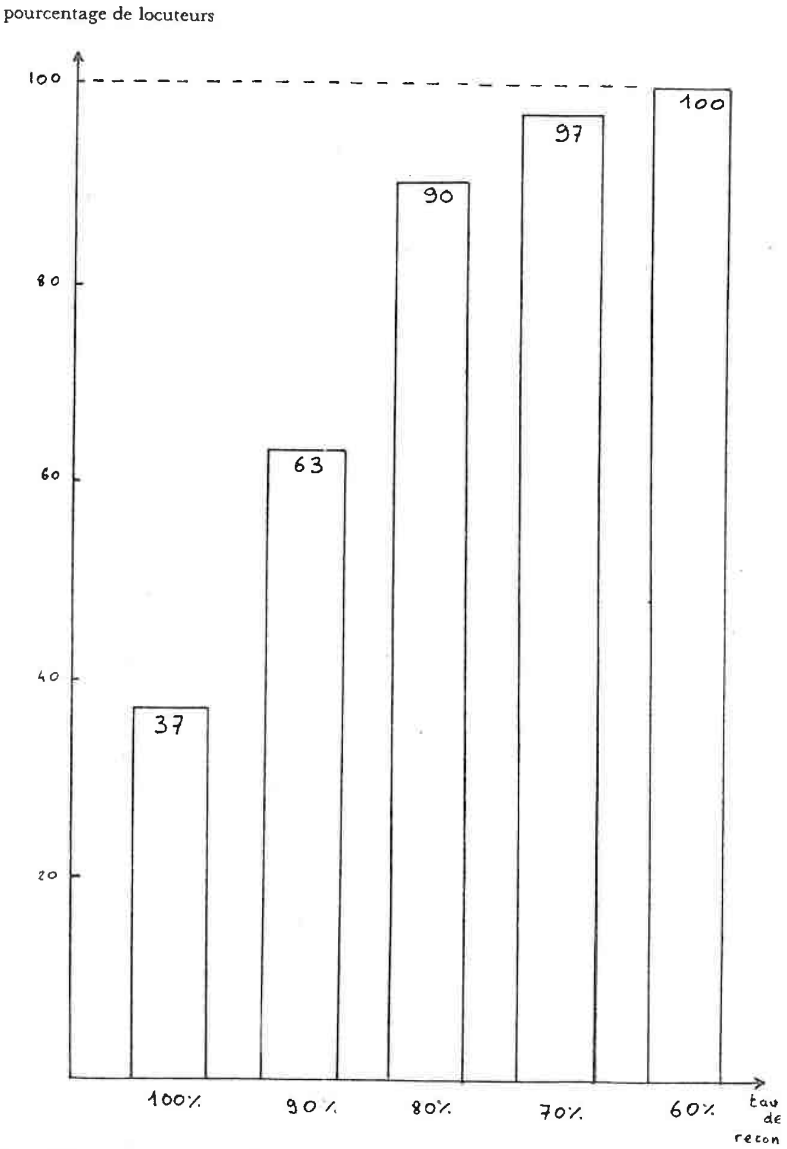


Figure B58: Courbe cumulée du pourcentage de locuteurs avec le taux de reconnaissance correspondant

73. Portée des conclusions

731. La classification

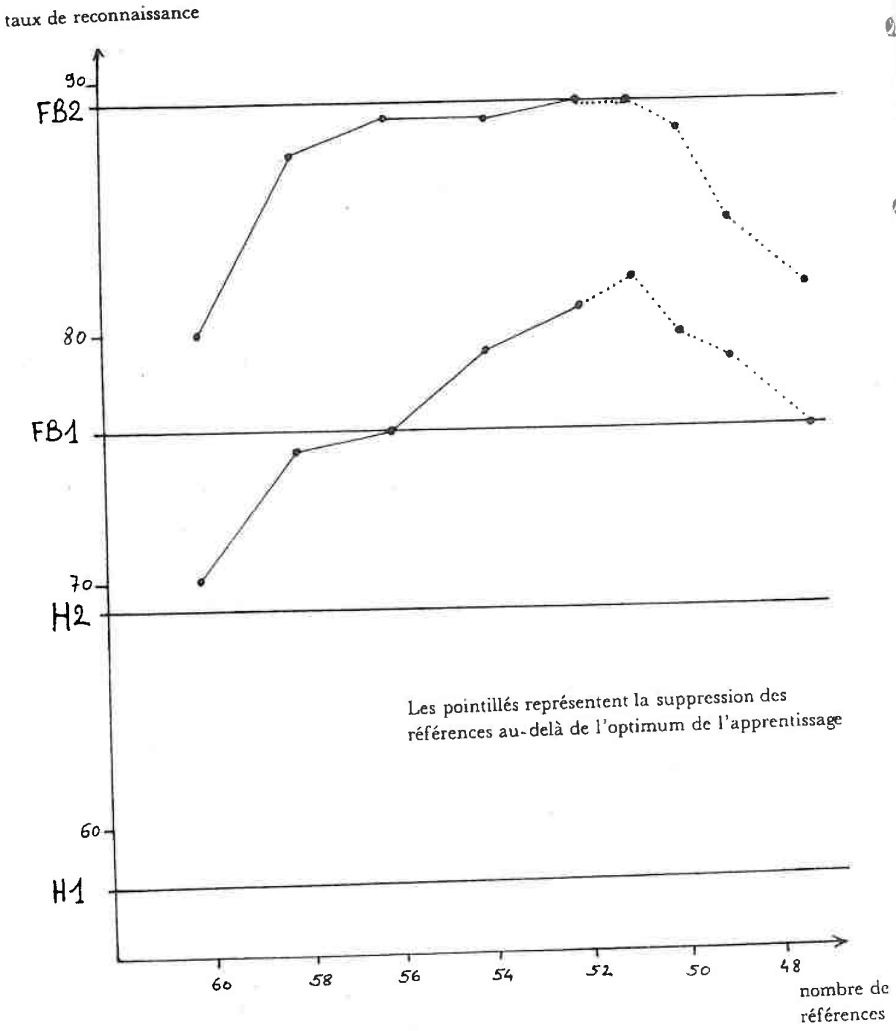
Le principe de base de ce travail était motivé par la quantité d'information mise en jeu dans la reconnaissance multilocuteur et consistait, dans le cadre de la reconnaissance globale de mots isolés, à conserver avec le minimum de prototypes, le maximum d'informations contenues dans le corpus d'apprentissage. On peut juger de la pertinence de l'étape de classification sur la figure B59. On a fait figurer deux seuils: Une ''borne inférieure '' obtenue par les résultats de reconnaissance sur 50 références choisies au hasard (5 références par type) et une ''borne supérieure '' obtenue par la ''force brute '' c'est-à-dire en utilisant la totalité du corpus d'apprentissage (1200 références). On constate qu'avec seulement 5% du volume de données initial, on conserve 91% de l'information brute et qu'après le traitement des points aberrants on dépasse même le taux de reconnaissance de la force brute (qui est assez faible, il est vrai, cf paragraphe précédent).

La méthode de CAH remplit donc bien son rôle de compression de l'information et on peut supposer qu'appliquée à un corpus de bonne qualité, elle donnerait de bons résultats. Rappelons que testée dans les mêmes conditions (corpus C20) que la méthode UWA des Bell Laboratories, elle donne de meilleurs résultats (cf paragraphe 633).

La pertinence de la classification peut également se constater dans le cas de la classification globale où les classes difficiles (classes empiètantes ou polymorphes, cf chapitre 5) sont surreprésentées par rapport aux classes homogènes et bien différenciées. On voit par exemple sur la figure B60a que les 50 références, issues d'une classification globale à segmentation guidée par le type, sont très inégalement réparties entre les différents types. Les locutions difficiles (sept, cinq, quatre) ont plus de prototypes que les autres et si on les visualise, on constate qu'elles couvrent bien les différentes prononciations possibles (par exemple pour ''sept'' : /set/, /se/, /et/). La classe ''un'' est également sursegmentée à cause du recouvrement qu'elle présente avec la classe ''cinq''.

Afin de mesurer la fiabilité relative de cette méthode, nous avons effectué des tests à partir du même corpus d'apprentissage sur le même corpus de reconnaissance (C60) mais en utilisant deux autres méthodes mises au point au laboratoire.

Toutes deux sont fondées sur un code-book obtenu par quantification vectorielle pour chaque mot du vocabulaire.



FB : Score obtenu avec la force brute (1200 références)
H : Score obtenu avec 50 références au hasard

Figure B59: Evolution des taux de reconnaissance en fonction du nombre de points aberrants supprimés en 1 et 2 propositions (corpus C60)

La première méthode utilise le corpus d'apprentissage pour construire pour chaque mot du vocabulaire, un automate d'états finis qui prenne en compte les différentes variations de prononciation qui ont été observées. La reconnaissance consiste ensuite en une comparaison dynamique entre le mot inconnu et chaque automate. cf [Smaïli 87].

La deuxième méthode utilise l'algorithme de Burton (cf Partie A, paragraphe 442b) comme préselection des candidats, puis détermine le meilleur à l'aide d'une comparaison dynamique à coefficients variables; le mot reconnu étant celui qui minimise la somme du score de comparaison dynamique et de la distorsion de codage (cf [Boyer 87]).

La figure B60b donne les résultats de reconnaissance en première proposition et par chiffre. Parmi les 3 méthodes testées, on constate que c'est la méthode de CAH qui donne les meilleurs résultats pour un volume de données comparable.

D'autres tests sont également envisagés afin de comparer ces résultats à ceux obtenus sur le même corpus avec une méthode fondée sur les réseaux markoviens (cf [Mari 85]). Les conclusions de ces tests feront l'objet d'un article à paraître.

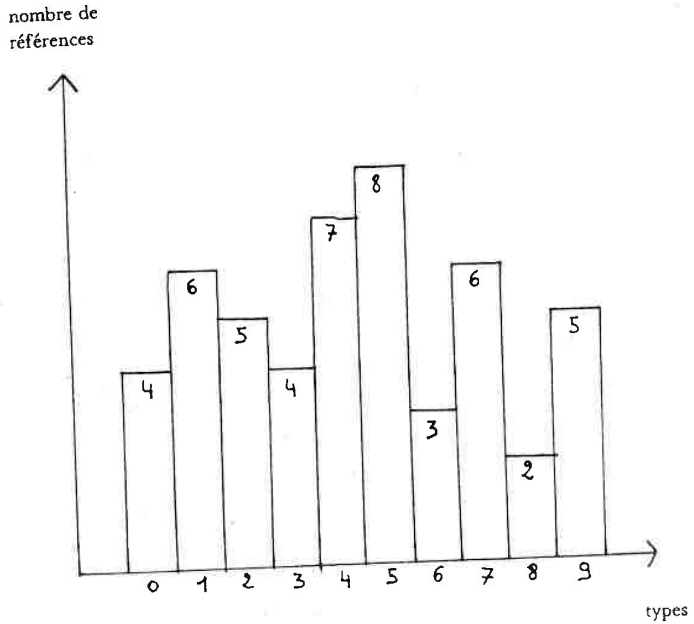


Figure B60a: Répartition du nombre de références par type dans la classification globale (C20)

	CAH	Methode 1	Methode 2
0	97	100	91
1	77	79	77
2	87	74	93
3	93	94	91
4	73	77	59
5	60	70	79
6	73	77	87
7	53	42	81
8	100	94	83
9	93	77	58
Tot	81	78	80

Figure B60b: Comparaison des taux de reconnaissance en 1 proposition suivant les méthodes utilisées

732. Choix des variantes

a) Critères stables

Le bien fondé de l'étape de classification et de la méthode utilisée étant acquis, il est intéressant de dégager un certain nombre de conclusions stables quant aux différents paramètres et variantes testés au cours de ce travail. Trois points essentiellement se sont révélés très stables lors des tests:

- Le critère d'agrégation optimum est la variance.(cf paragraphe 634).
- Le meilleur choix de représentant d'une classe est le point minimax de cette classe (cf paragraphe 64).
- Le traitement des points aberrants s'est toujours montré efficace (voir paragraphe c).

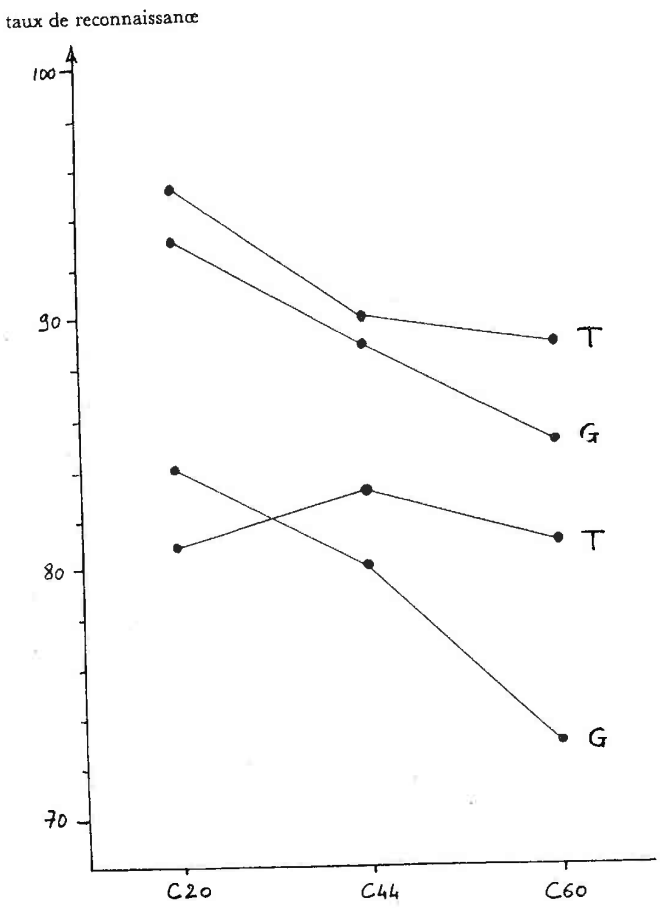
b) Choix du corpus

Pour ce qui est du choix du corpus à classer, il semble que l'avantage de la classification globale par rapport à la classification par type s'estompe lorsque croît le nombre de locutions de l'apprentissage. Sur un petit corpus (C20), les taux de reconnaissance prouvent nettement la supériorité de la classification globale associée à la prise en compte des rayons lors de la reconnaissance, le critère optimal étant alors l'intériorité.

Sur le corpus C44 les deux méthodes donnent des résultats comparables, et sur le corpus C60 la classification globale effectuée à partir des résultats de la classification par type donne de moins bons résultats à nombre de références équivalent.

La figure B61 confirme que la classification par type est moins sensible à l'augmentation du nombre de locuteurs; on réservera donc la classification globale avec prise en compte des rayons pour les petits corpus. Elle s'avère en particulier intéressante pour la reconnaissance plurilocuteur jusqu'à des corpus de 50 locuteurs (cf figure B56). On pourra donc l'utiliser dans les cas d'applications où les utilisateurs sont tous connus lors de l'apprentissage, elle donne dans ce cas des taux de reconnaissance en 2 propositions proches de 100%.

La classification par type au contraire sera réservée pour les corpus plus étendus où, de toutes façons, elle demeure la seule possible. Les tests montrent qu'il n'est pas souhaitable, dans ce cas, de tenir compte des rayons. Ce qui tend à prouver que l'avantage qu'apportent les rayons lors de la reconnaissance globale tient à l'information qu'ils contiennent sur le



T : Classification par type

G : Classification globale

Figure B61· Comparaison des classifications globale et par type en 1 et 2 propositions selon les trois corpus utilisés

recouvrement des classes de types différents.

Le critère de segmentation le plus simple et le plus efficace est le critère du nombre de classes fixé à l'avance. C'est un paramètre dont dépend le temps de reconnaissance donc qu'il est intéressant de pouvoir controler en fonction des applications souhaitées.

Remarquons une caractéristique très stable: Il vaut toujours mieux choisir pour la classification plus de classes et conserver seulement les N plus grosses. Ainsi, par exemple, on gagne 5 points de reconnaissance en effectuant d'abord une classification en 10 classes par type et en sélectionnant les 6 plus grosses, par rapport à une classification qui donne directement 6 classes par type. Les 4 classes ainsi éliminées regroupant peu de points, on peut supposer qu'elles correspondent à des locutions atypiques qui se révèlent néfastes lors de la reconnaissance. Ceci confirme la constatation faite par [Briant 84] pour le système Syril.

c) Traitement des points aberrants.

Dans tous les cas testés et quel que soit le corpus, il a toujours été avantageux de traiter les points aberrants (c'est-à-dire des références qui perturbent la reconnaissance). Il s'est d'ailleurs souvent révélé plus intéressant de supprimer de cette manière des références que d'en ajouter. Le traitement des points aberrants s'effectue à plusieurs niveaux:

- Lors de la classification globale par l'application d'un coefficient d'homogénéité des classes (coefhomo) qui permet d'éliminer, dans une classe majoritairement d'un certain type, les quelques points d'un type différent. (cf paragraphe 635e).
- Lors de la classification par type par l'élimination des classes de cardinal faible (cf paragraphe précédent).
- Dans tous les cas par le traitement des ''singletons'' (cf paragraphe 462).

Il convient de remarquer que ce traitement a été étendu à tous les points aberrants, qu'ils correspondent ou non à des points isolés de la classification. En effet, on constate principalement sur le corpus C60, que certaines références sont à la fois utiles (c'est-à-dire souvent utilisées pour une reconnaissance de leur type) et néfastes (c'est-à-dire utilisées pour une reconnaissance d'un type différent).

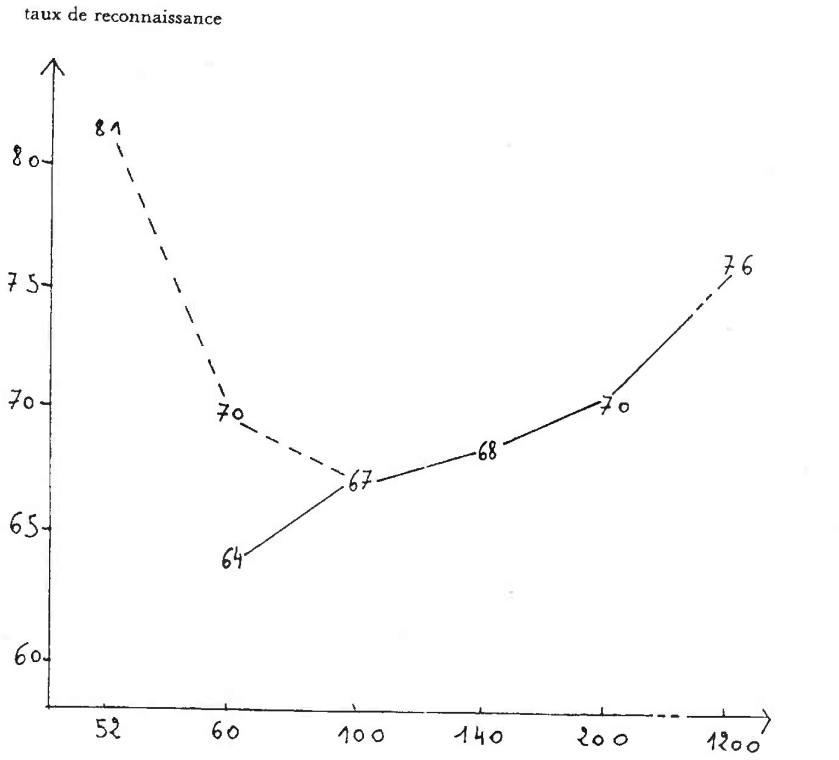
Certaines d'entre elles, peu nombreuses, sont présentes lors de la reconnaissance de presque tous les types et de nombreuses fois. Il serait donc avantageux de supprimer ces prototypes car on éliminerait ainsi de fréquentes confusions et on constate que d'autres références prennent le plus souvent leur place pour une reconnaissance du même type. Ces points aberrants (que j'ai appelés ''pollueurs''), sont donc d'une autre nature que

les points isolés déjà vus, mais on peut faire à leur sujet la même remarque: Les prototypes ''pollueurs'' et les chiffres ''pollués'' sont toujours les même à l'apprentissage et à la reconnaissance, ce qui justifie pour eux aussi le traitement appliqué aux singletons (cf paragraphe 462). On remarque sur la figure B59, l'évolution des taux de reconnaissance au fur et à mesure des points pollueurs supprimés; là encore les optima de la reconnaissance et de l'apprentissage coincident presque exactement: L'optimum de l'apprentissage a lieu pour 52 références, celui de la reconnaissance pour 51 références. Si l'on continuait de supprimer les références après l'optimum, on gagnerait donc un point de reconnaissance mais ensuite les taux chutent rapidement.

On constate également qu'il est en général plus efficace de supprimer les prototypes pollueurs que d'augmenter le nombre des références, on voit en effet sur la figure B62 que le taux de reconnaissance évolue assez lentement avec le nombre de références brutes (courbe en trait plein), beaucoup moins vite en tous cas qu'avec la suppression des prototypes correspondant aux classes les plus petites et des pollueurs (la courbe en pointillé en est un exemple à partir de 10 références par type).

733. Le nombre de locuteurs

Les mêmes méthodes et algorithmes ont été testés sur trois corpus de tailles différentes (20, 44 et 60 locuteurs). La figure B63 montre l'évolution du nombre de références optimal et du taux de compression de l'information correspondant, en fonction de cette taille. On constate que le nombre de références à conserver pour la reconnaissance augmente beaucoup moins vite que le nombre de locutions: Pour un corpus 12 fois plus important on n'a multiplié la quantité d'information que par 1.2. Le nombre de références optimal semble se situer dans ce contexte entre 4 et 6 références par type de mot et représente pour les corpus les plus importants, plus de 80% de l'information pour 5% du volume des données.



La courbe en pointillé représente le traitement des points aberrants

Figure B62: Evolution des taux de reconnaissance en fonction du nombre de références (corpus C60)

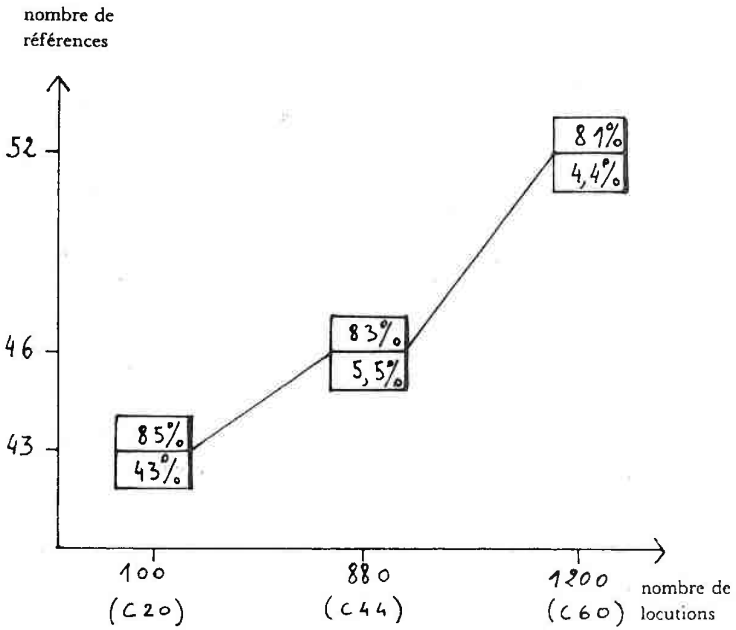


Figure B63: Evolution du taux de compression de l'information avec le nombre de locutions du corpus d'apprentissage

CHAPITRE 8

CONCLUSION

Dans cette étude nous avons abordé le problème de la reconnaissance de mots multilocuteur. Nous nous sommes limités dans un premier temps à un contexte assez simple (mots isolés, petit vocabulaire) mais suffisant pour bon nombre d'applications; Le système MIMULE fonctionne actuellement sur une simulation de reconnaissance interactive de codes chiffrés (applications bancaires, numéros de téléphone...).

Notre but était de prouver la validité, dans ce contexte, d'une méthode de réduction de données simple et réputée efficace : la classification ascendante hiérarchique (CAH) que nous avons appliquée aux locutions entières.

Nous nous sommes attachés à mettre en évidence les meilleurs paramètres permettant de rendre compte de la variabilité multilocuteur, à partir d'une structure de base constante:

- Comparaison dynamique globale des locutions d'apprentissage,
- Classification de ces locutions à partir de la matrice de leurs pseudo-distances,
- Extraction d'un représentant par classe,
- Reconnaissance par comparaison dynamique globale aux représentants et assimilation au plus proche voisin.

Cette méthode s'avère donner de bons résultats compte tenu de la difficulté du corpus utilisé et des paramètres imposés ne dépendant pas de cette application (acquisition, segmentation et algorithme de comparaison dynamique). Les principales conclusions auxquelles nous avons abouti sont les suivantes:

1) Comparée à d'autres méthodes sur le même corpus (classification UWA, automates, quantification vectorielle plus méthode de Burton) la CAH aboutit à des taux de reconnaissance supérieurs pour un volume de données tout à fait acceptable: Avec le corpus d'apprentissage de 1200

locutions, on conserve plus de 80% de l'information avec seulement 5% de la quantité de données.

2) Lors d'une CAH le meilleur critère d'agrégation est la variance, et le meilleur choix de centroide est le point minimax.

3) Suivant le contexte de l'application qui est visé, deux variantes de la CAH peuvent être utilisées:

* En contexte plurilocuteur (nombre limité d'utilisateurs, tous connus lors de l'apprentissage), la méthode optimale est la suivante:
  - Classification globale tous types confondus,
  - Segmentation de l'arbre hiérarchique guidée par le type des feuilles,
  - Reconnaissance avec prise en compte des rayons des classes (critère de l'intériorité).

* En contexte multilocuteur vrai (les utilisateurs du système sont différents de ceux qui ont participé à l'apprentissage), la méthode optimale est la suivante:
  - Classification par type séparément,
  - Segmentation guidée par le nombre de classes et conservation des prototypes correspondant aux classes les plus volumineuses (6 références par type permettent de bons résultats),
  - Reconnaissance par le plus proche voisin.

4) Dans tous les cas, il est avantageux de faire subir aux références sélectionnées, un traitement en vue d'éliminer un certain nombre de points isolés et de "pollueurs" qui se révèlent plus néfastes qu'utiles lors de la reconnaissance.

Les différents tests de pertinence de la classification (comparaison des méthodes entre elles et par rapport à la force brute), ont montré que l'étape de sélection des prototypes par classification est probablement arrivée à un point proche de son optimum et que l'amélioration des taux de reconnaissance bute sur une limite due à la non adéquation de la notion de distance utilisée, avec le concept humain.

Pour améliorer l'efficacité de la reconnaissance, nous devrons donc plutôt nous tourner vers une amélioration des processus d'acquisition (détection des frontières de mot, extraction des paramètres) et vers un algorithme de comparaison qui respecte mieux la notion acoustique humaine de "différence entre les locutions".

Nous envisageons également une autre application de ce travail associé à la méthode décrite dans [Boyer 87] (quantification vectorielle et comparaison dynamique) consistant à utiliser l'algorithme de CAH à deux niveaux:

- Une classification hiérarchique des vecteurs de parole du corpus d'apprentissage aboutit à un code-book conservant la structure arborescente (calquée sur la méthode CSA du paragraphe 67). Lors de la reconnaissance, le codage de la forme inconnue est accéléré par une recherche dichotomique dans l'arbre hiérarchique obtenu.
- Sur les locutions d'apprentissage ainsi quantifiées, on effectue ensuite une seconde classification hiérarchique avec segmentation afin d'obtenir les ptototypes les plus représentatifs

Lors de la reconnaissance, une étape rapide de préselection fondée sur la distorsion de codage minimum, conserve seulement quelques candidats. Le mot inconnu codé est comparé par programmation dynamique aux prototypes conservés; la décision finale est prise en optimisant une fonction de la distorsion et du score de comparaison.

Cette application est en cours de développement sur un vocabulaire assez étendu mais bien différencié : Il s'agit, dans le cadre d'un centre de renseignements SNCF, de traiter un vocabulaire constitué de 100 noms de villes françaises plus quelques mots-clé, la reconnaissance intervenant lors d'un dialogue fortement dirigé par l'ordinateur.

Une autre extension prévue de ces méthodes consiste à effectuer une classification sommaire des locuteurs (par exemple sur les phonèmes i-a-ou), puis à créer un code-book par type de locuteur ainsi déterminé. La reconnaissance passe alors par une courte phase d'adaptation au nouveau locuteur permettant de déterminer son type et donc d'obtenir le code-book le plus adéquat.

Nous envisageons également de généraliser la méthode de présélection de Burton à la reconnaissance multilocuteur de grands vocabulaires. Pour celà, nous proposons au cours de l'apprentissage, de hiérarchiser les code-books de chaque mot et d'éliminer, lors de la recherche dichotomique, les branches correspondant à des distorsions trop importantes. Cette présélection permet d'éviter un grand nombre de comparaisons inutiles avec des mots très différents du mot prononcé et d'accéder rapidement à une classe de candidats. La décision finale doit ensuite faire intervenir un critère temporel (comparaison dynamique ou autre) afin de choisir et de proposer le candidat optimal.

BIBLIOGRAPHIE

[Andreewsky 85]: Andreewsky A., Desi M., Poirier F., 'SHERPA: De l'étiquetage phonétique automatique à la reconnaissance par analyse ternaire ', 5eme congrès AFCET RFIA, Nov 1985.

[Arcella 86]: Arcella F., Fissore L., Oreglia M., Pirani G., ' speaker independant recognition of isolated words: Template clusters and Markov models', ICPR, Paris,1986.

[Bahl 79]: Bahl L.R., Baker J.K., Cohen P.S., Cole A.G., Jelinek F., Davies B.L., Mercer R.L., ' recognition results with several experimental acoustic processors', IEEE ICASSP, Washington, 1979.

[Ball 65]: Ball G.H., Hall D.J., 'Isodata, a novel method of data analysis and pattern classification', Technical report of Stanford Research Institute, 1965.

[Baker 81]: Baker J., 'How to achieve recognition: A tutorial status report on automatic speech recognition', speech technology, 1(1), 1981.

[Baum 72]: Baum , Leonard E. 'An inequality and associated maximisation technique in statistical estimation for probabilistic functions of a Markov process', Inequalities, vol3, 1972.

[Bellman 57]: Bellman J., 'Dynamic programming', Princeton University Press, 1957.

[Bourlard 84]: Bourlard H., Ney H., Wellekens C.J., 'Connected digits recognition using vector quantization' IEEE ICASSP, San Diego, 1984.

[Bourlard 87]: Bourlard H., Wellekens C.J., 'Speech pattern discrimination and multilayer perceptrons' Manuscrit M211, Philips-MBLE, Sept 1987.

[Boyer 87]: Boyer A. 'Application des techniques de programmation dynamique et de quantification vectorielle à la reconnaissance des mots isolés et des mots enchaînés', thèse de doctorat d'université, Université de Nancy I, 1987.

[Briant 84]: Briant N., Flocon B., 'Syril: Système temps réel de reconnaissance de mots indépendants du locuteur' 4eme congrès AFCET RFIA, 1984.

[Bridle 82]: Bridle J.S., Brown MD., Chamberlain R.M., 'An algorithm for connected word recognition',IEEE ICASSP, pp.899-902, Paris, 1982.

[Bridle 83]: Bridle J.S., Chamberlain R.M., 'Automatic labelling of speech using synthesis by rule and non linear time alignment', Elsevier Sciences Publishers B.V. North Holland, 1983.

[Broad 74]: Broad D.J., Shoup J.E. 'Concepts for acoustic phonetic recognition', Speech recognition IEEE Symposium, 1974. Academy Press 1975.

[Burton 83]: Burton D.K., Shore J.E., Buck J.T., 'A generalization of isolated word recognition using vector quantization',IEEE ICASSP, Boston,pp. 1021-1024, 1983.

[Burton 85]: Burton D.K., 'Applying matrix quantization to isolated word recognition', ICASSP, Tampa, pp.29-32, 1985.

[Buzo 80]: Buzo A.,Gray A.H., Gray R.M., Markel J.D., 'Speech coding based upon vector quantization',IEEE Trans. ASSP, vol28, no 5, 1980.

[Carton 74]: Carton F., 'Introduction à la phonétique du français', Bordas Etudes No 303, 1974.

[Chandon 81]: Chandon J.L., Pinson S., 'Analyse typologique. Théorie et applications', Ed. Masson, 1981.

[Charpillet 85]: Charpillet F., 'Un système de reconnaissance de parole continue pour la saisie de textes lus', thèse d'université, Université de Nancy I, 1985.

[Cravero 84]: Cravero M., Fissore L., Pieraccini R., Scagliola C., 'Syntax driven recognition of connected words by Markov models' IEEE ICASSP,San Diego, 1984.

[Damestoy 86]: Damestoy J.P., 'Réalisation d'un système à base de

prototypes pour le controle du décodage acoustico-phonétique de la parolé, thèse de 3eme cycle, Université de Nancy I, 1986.

[Diday 70]: Diday E. 'Une nouvelle méthode en classification automatique et reconnaissance des formes: La méthcde des nuées dynamiques', Revue de statistiques appliquées, 19(2), 1970.

[Divoux 82]: Divoux P. 'Reconnaissance de mots isolés multilocuteurs par classification statistique ', Rapport de DEA, Université de Nancy I, 1982.

[Divoux 85]: Divoux P., Haton J.P., 'Classification hiérarchique et reconnaissance multilocuteur',5ème Congrès AFCET RFIA, 1985.

[Duda 66]: Duda R.O., Fossum H., 'Pattern classification by iteratively determined linear and piecewise linear discriminant functions', IEEE Trans. on Ele. Comp., 1966.

[Feldman 84]: Feldman J.A., Gaverick S.L., Rhodes F.M., Mann J.R. 'A wafer scale integration systolic processor connected words recognition', IEEE ICASSP, 1984.

[Flanagan 65]: Flanagan J.L. 'Speech analysis, synthesis and perception', Springer Verlag, 1965.

[Flocon 86]: Flocon B, Lockwood P., Sap J, Sauter L., 'MARIPA : speaker independant recognition of speech on IBM-PC', ICPR, Paris, 1986.

[Fohr 86]: Fohr D., 'Aphodex: un système expert en décodage acoustico-phonétique de la parole continue ', Thèse de l'Université de Nancy I, 1986.

[Fonsale 84]: Fonsale P. 'Simulation informatique d'un système multilocuteur de reconnaissance de la parole (mots isolés) sans apprentissage oral. Analyse par traits phonétiques', Thèse de docteur ingénieur, INPG, Grenoble, 1984.

[Fonsale 84b]: Fonsale P. 'Feature based speaker independant recognition without oral learning', IEEE, ICASSP, 1984.

[Frison 83]: Frison P. 'Un processeur intégré pour la reconnaissance de la parolé, INRIA research report, N.215, July 1983.

[Fu 74]: Fu K.S., 'Syntactic methods in pattern recognition', Academic

press, 1974.

[Fukunaga 73]: Fukunaga K., 'An introduction to statistical pattern recognition', Academic press, 1973.

[Fukunaga 75]: Fukunaga K., Narendra P., 'A branch and bound algorithm for computing K-nearest neighbors', IEEE Trans. on computers, 1975.

[Furui 80]: Furui S., 'A training procedure for isolated word recognition systèms', IEEE Trans. ASSP, vol28, no 2, 1980.

[Gagnoulet 84]: Gagnoulet C., Jouvet D., 'Seraphine: système de reconnaissance de courtes phrases', 13eme JEP, Bruxelles,1984.

[Gertsman 86]: Gertsman L.J. ' Classification of self normalized vowel',IEEE Trans. AU, vol16, no 1, 1986.

[Gillet 84]: Gillet & all, 'SERAC, Un système expert en reconnaissance acoustico-phonétique ', 4eme congrès AFCET RFIA, Paris, 1984.

[Gourinda 87]: Gourinda A., Haton J.P. 'Reconnaissance rapide de mots isolés par quantification vectorielle multisection', 16eme 13eme JEP, Hammamet, 1987.

[Gupta 78]: Gupta V.N., Bryan J.K., Gowdy J.N., 'Speaker-independant vowel identification in continuous speech', IEEE ICASSP, Tulsa, 1978.

[Gupta 82]: Gupta V., Mermelstein P., 'Effects of speaker accent on the performance of a speaker-independant, isolated word recognizer', JASA 71(6), June 1982.

[Gupta 84]: Gupta V., Lennig M., Mermelstein P., 'decision rules for speaker independant isolated word recognition', IEEE ICASSP, 1984.

[Haton 82]: Haton J.P., 'Recherches et développements actuels en reconnaissance automatique de la parole ', 4eme journée francophone sur l'informatique, Geneve, 1982.

[Haton 85]: Haton J.P., 'Intelligence artificielle en compréhension automatique de la parole ', TSI, vol4, no3, 1985.

[Haton MC 85]: Haton M.C., 'Contribution à l'éducation vocale assistée par ordinateur: étude des voix et réalisation du système SIRENE ', Thèse

d'état, Université de Nancy I, 1985.

[Hoskins 85]: Hoskins J., 'Large vocabulary speech recognition, today at IBM', Speech technology, Aug. Sept 1985.

[Jakobson 63]: Jakobson R., Fant G. et Halle M., 'Preliminaries to speech analysis'. Cambridge MA, MIT Press, 1963.

[Jambu 78]: Jambu M., Lebeaux M.O. ' Classification automatique pour l'analyse des données', Dunod, 1978.

[Jelinek 76]: Jelinek F, 'Continuous speech recognition by statistical methods', Proc. IEEE vol 64 no 4, avril 1976.

[Jelinek 87]: Jelinek F, 'Experiments with the Tangora 20000 words speech recognizer' IEEE ICASSP, 1987.

[Johanssen 83]: Johanssen J.,Mac Allister J., Michael T., Ross S. 'A speech spectrogram expert', IEEE ICASSP, Boston, 1983.

[Jouvet 86]: Jouvet D., Monne J., Dubois D. 'A Network-based speaker independant connected words recognition system IEEE ICASSP, 1986.

[Juang B.H., Rabiner L.R., Levinson S.E., Sondhi M.M., 'Recent developments in the application of hidden Markov models to speaker independant isolated words recognition', IEEE ICASSP, 1985.

[Kamgar-Parsi 85]: Kamgar-Parsi B., Kanal L.N., 'An improved branch and bound algorithm for computing K-nearest neighbors', Pattern recognition letters 3, Jan 1985.

[Kasuya 82] : Speech technology, Aug. Sept 1985.

[Lecorre 79]: Lecorre C., Vives R. 'Un programme de cadrage pour l'adaptation au locuteur en reconnaissance automatique de la parole ',3eme congrès AFCET RFIA,1979.

[Lelievre 81]: Lelièvre A. 'Codage et traitement de certains types de données phonétiques pour une reconnaissance automatique de la parole par classification ', thèse de 3eme cycle , Université de Rennes I, 1981.

[Lennig 83]: Lennig M. 'Automatic alignment of natural speech with a corresponding transcription', Elsevier Sciences Publishers B.V. North Holland 1983.

[Le Ny 80]: Le Ny J.F., Denis M., 'Identification et compréhension du langage naturel: Perspectives cognitives', GALF AFCET 1980.

[Levinson 79]: Levinson S.E., Rabiner L.R., Rosenberg S.E., Wilpon J.G., 'Interactive clustering techniques for selecting speaker-independant reference templates for isolated word recognition',IEEE Trans. ASSP, vol27, no 2, 1979.

[Levinson 83]: Levinson S.E., Rabiner L.R., Sondhi M.M., ' Speaker independant isolated digit recognition using Markov models', IEEE ICASSP, Boston, 1983.

[Lienard 72]: Lienard J.S., 'Analyse, synthèse et reconnaissance automatique de la parolé Thèse d'etat, Université de Paris VI, 1972.

[Lienard 82]: Lienard J.S., Mariani J.J., 'Système de reconnaissance acoustique de mots isolés: Moise ', Dossier ANVAR N 50312, juin 1982.

[Linde 80]: Linde Y., Buzo A., Gray R.M., 'An algorithm for vector quantizer design', IEEE Trans. Commun. Technol., vol COM-28, janvier 1980.

[Lloyd 57]: Lloyd S.P., 'Least squares quantization in PCM' (1957), IEEE Trans. IT, vol 28, mars 1982.

[Lockwood 84]: Lockwood P., 'Proposition de mots dans un système de reconnaissance de mots isolés multilocuteurs: une approche en vue du traitement des grands vocabulaires', 13eme JEP, Bruxelles 1984.

[Lockwood 85]: Lockwood P., 'Accès rapide dans un dictionnaire de mots', 5eme Congrès AFCET RFIA, Nov 1985.

[Lockwood 86]: Lockwood P., 'A low cost DTW-based discrete utterance recogniser', IEEE ICASSP, 1986.

[Lowerre 77]: Lowerre B.T. 'Dynamic speaker adaptation in the Harpy speech understanding systèm', IEEE ICASSP, Hartford, 1977.

[Mac Queen 67]: Mac Queen J., 'Some methods for classification and analysis of multivariable observations', 5th Berkeley symposium on mathematics, statistics and probability, 1(1), 1967.

[Makhoul 72]: Makhoul J.I., WOLF J.J. 'Linear prediction and the spectral analysis of speech', Bolt, Beranek & Newman Inc. Report 2304,

Cambridge MA, 1972.

[Makino 84]: Makino S., Kido K., 'A speaker independant word recognition system based on phoneme recognition for a large size vocabulary', IEEE ICASSP, 1984.

[Malmberg 71]: Malmberg G. 'La phonétique ', Que sais-je ?, PUF, Paris, 1971.

[Mari 84]: Mari J.F., Haton J.P., 'Some experiments in automatic recognition of a thousand words vocabulary', IEEE ICASSP, San Diego, 1984.

[Mari 85]: Mari J.F., 'Reconnaissance de mots enchainés à l'aide de modèles markoviens discrets', 5 ème congrès AFCET-RFIA, pp 859-867, Grenoble, 1985.

[Mariani 84]: Mariani J., Eskenazi M., Majani E., Carrier A., Defrance R., 'Expériences en reconnaissance de mots isolés multilocuteur multiréférencé, 13eme JEP, Bruxelles 1984.

[Mariani 85]: Mariani J., 'Speech technologies in western Europe, a review', Speech technology, Aug. Sept 1985.

[di Martino 84]: di Martino J., 'Contribution à la reconnaissance globale de la parole: mots isolés et mots enchainés', Thèse de docteur ingénieur en informatique, Université de Nancy I, 1984.

[Memmi 84]: Memmi D., Eskenazi M., Mariani J., Stern P., 'SONEX: Système expert en lecture de spectrogrammes', Rapport ATP IA, 1984.

[Messaoudi 86]: Messaoudi A., Gauvain J.L. ' reconnaissance multilocuteur de mots isolés fondée sur une approche phonétique ',15eme JEP, Aix en Provence, 1986.

[Miclet 83]: Miclet L., Dabouz M., 'Approximative fast nearest neighbour recognition', Pattern recognition letters 1, july 1983.

[Miclet 84]: Miclet L., Dabouz M., 'Un vocodeur à classification: transmission de la parole à très faible débit par quantification vectorielle du spectre ', ENST , 1984.

[Miclet 84]: Miclet L., 'Méthodes structurelles pour la rec des formes', Eyrolles 1984.

[Miclet 85]: Miclet L., Faure C., 'Reconnaissance des formes structurelles: Developpement et tendances', 5eme congrès AFCET RFIA, nov 1985.

[Moreau 84]: Moreau J.V., 'Inférence en classification hiérarchiqué, Thèse de docteur ingénieur, ENST, 1984.

[Morii 85]: Morii S., Niyada K., Fujii S., Hoshimi M., 'Large vocabulary speaker independant japanese recognition system',IEEE ICASSP, 1985.

[Mokeddem 85]: Mokeddem A., Hugli H., Pellandini F., 'Reconnaissance multilocuteur de mots isolés', 14eme JEP, Paris, 1985.

[Neel 83]: Neel F., Eskenazi M., Mariani J.J. 'Cadrage automatique pour la constitution de dictionnaire d'entités phonétiques', Elsevier Sciences Publishers B.V. North Holland, 1983.

[Niles 83]: Niles L., Silverman H.F., Dixon N.R., 'A comparison of three feature vector clustering procedures in a speech recognition paradigm',IEEE ICASSP, Boston, 1983.

[Pan 85]: Pan K., Soong F., Rabiner L., 'A vector quantization based preprocessor for speaker-independant isolated word recognition', IEEE ICASSP, June 1985.

[Peterson 52]: Peterson G.E., Barney H.L., 'Control methods used in a study of the vowels', JASA 24(2), 1981.

[Pierrel 81]: Pierrel J.M., 'Etude et mise en ouvre de contraintes linguistiques en compréhension automatique du discours continu', Thèse d'etat, Université de Nancy I, 1981.

[Pister 84]: Pister C., 'Adaptation au locuteur par apprentissage automatique, application à un système de reconnaissance automatique de la parole ', thèse de 3eme cycle, Université de Nancy I, 1984.

[Pollard 82]: Pollard D.V., 'Quantization and the method of the K-means', IEEE Trans. I.T., vol28, no 2, 1982.

[Rabiner 76]: Rabiner L.R., Cheng M.J., Rosenberg A.E., Mac Gonegal C.A., 'Comparative performance study of several pitch detection algorithms', IEEE Trans. ASSP, vol24, 1976.

[Rabiner 79a]: Rabiner L.R., Levinson S.E, Rosenberg A.E., Wilpon J.G., ' speaker-independant recognition of isolated word using clustering

techniques', IEEE Trans. ASSP, vol27, no 4, 1979.

[Rabiner 79b]: Rabiner L.R., Wilpon J.G., 'Considerations in applying clustering technics to speaker-independant word recognition', JASA 66(3), Sept 1979.

[Rabiner 84]:Rabiner L.R., Levinson S.E., Sondhi M.M., 'On the use of hidden Markov models for speaker-independant recognition of isolated words from a medium size vocabulary', BSTJ 63(4), April 1984.

[Rossi 83]: Rossi M., Nishinuma Y., Mercier G., 'Indices acoustiques multilocuteurs et independants du contexte pour la reconnaissance automatique de la parole ', Elsevier Sciences Publishers B.V. North Holland, 1983.

[Sakoe 71]: Sakoe H., Chiba S., 'A dynamic programming approach to continuous speech recognition', Proc. International Congress on Acoustics, Budapest, 1971.

[Sakoe 78]: Sakoe H., Chiba S., 'Dynamic programming algorithms optimization for spoken word recognition', IEEE Trans. ASSP, vol26, no 1, 1978.

[Schafer 74]: Schafer R.W., Rabiner L.R., 'Parametric representation of speech', Speech recognition, IEEE symp 1974.

[Schinke 86]: Schinke D., ' Speaker independant recognition applied to telephone access information systems', Speech tech' 1986.

[Shore 83]: Shore J.E., Burton D.K., 'Discrete utterance speech recognition without time alignment', IEEE Trans. IT, vol29, no 4, 1983.

[Simon 84]: Simon J.C. 'La reconnaissance des formes par algorithme ', Masson, 1984.

[Smaili 87]: Smaili K. 'Realisation d'une machine multiniveaux pour la reconnaissance de la parole continue ' Rapport de DEA, Université de Nancy I, 1897.

[Stevens 71]: Stevens K.N., 'Source of inter and intra speakers variability in the acoustic properties of speech sounds', 7eme Int. Cong. Phon. Sci., 1971.

[Sundberg 80]: Sundberg J., 'Sons et musique ', Bibliothèque pour la

science, 1980.

[Sugamura 83]: Sugamura N., Shikano K., Furui S., 'Isolated words recognition using phonem-like templates', IEEE ICASSP, Boston, 1983.

[Tassy 84]: Tassy A., 'Chaine de Markov dans les systèmes de reconnaissance de la parole, étude bibliographiqué. MATRA-ENST, Mars 1984.

[Tassy 85]: Tassy A., Miclet L., ' Quantification vectorielle et reconnaissance de mots multilocuteur', 5eme congrès AFCET, Grenoble 1985.

[Tassy 86]: Tassy A., Miclet L., 'Reconnaissance multilocuteur des chiffres français par association d'un prétraitement fondé sur la QV avec des modèles de Markov cachés', pp 251-254, 15 ème JEP, Aix en Provence, mai 1986.

[Vidal-Ruiz 85]: Vidal-Ruiz E, Casacuberta-Nolla F, Rulot-Segovia H, 'Is really the DTW distance a metric?' Speech Communication, vol4, no 4, 1985.

[Wagner 81]: Wagner M., ' Automatic labelling of continuous speech with a given phonetic transcription using dynamic programming algorithms', IEEE ICASSP, Boston, 1981.

[Wakita 77]: Wakita H. 'Normalization of vowels by vocal tract length and its application to vowel identification', IEEE Trans. ASSP, vol25, 1977.

[Weinstein 75]: Weinstein C.J., Candless S.S., Mondshein L.S., Zue V.W., 'A system for acoustic-phonetic analysis of continuous speech',IEEE Trans. ASSP, vol23, 1975.

[Weste 83]: Weste N., Burr D.J., Ackland B.D. 'Dynamic time warp pattern matching using an integrated multiprocessing array. IEEE Trans. Computers, vol32, no 8, 1983.

[Williams 70]: Williams C.E., Stevens K.N., Hecker M.H.L., 'Acoustic manifestation of emotional speech', JASA 47, 66(a), 1970.

[Wilpon 82]: Wilpon J.G. Rabiner L.R., Bergh A., 'Speaker-independant isolated word recognition using a 129 words airline vocabulary', JASA 72(2), 1982.

[Wilpon 85]: Wilpon J.G. Rabiner L.R., 'A modified K-means algorithm for use in isolated word recognition', IEEE ICASSP, 1985.

[Zue 82]: Zue V.W., 'Acoustic phonetic knowledge representation: Implications for spectrogram readings experiments'. Automatic speech analysis and recognition, J.P. Haton Ed., D. Reidel, 1982.

NOM DE L'ETUDIANT : DIVOUX Pascal

NATURE DE LA THESE : Doctorat de l'Université de NANCY I en Informatique

VU, APPROUVE ET PERMIS D'IMPRIMER

NANCY, le 43

LE PRESIDENT DE L'UNIVERSITE DE NANCY I

RESUME

La difficulté principale des systèmes de reconnaissance automatique de la parole est due à la grande variabilité du signal vocal; celle-ci culmine dans les systèmes indépendants du locuteur.

Nous présentons ici un tel système appliqué à un contexte restreint: mots isolés, vocabulaire limité, comparaison dynamique globale.

Lors de la phase d'apprentissage du système, les références à conserver sont sélectionnées par des méthodes statistiques de réduction de données. Plusieurs algorithmes ont été testés et pour chacun d'eux nous avons tenté de déterminer les paramètres optimaux avec leur influence sur les temps de réponse, l'encombrement mémoire et les taux de reconnaissance. Le corpus utilisé est constitué de 60 locuteurs (30 hommes et 30 femmes), moitié pour l'apprentissage, moitié pour les tests de reconnaissance.

L'algorithme optimal s'avère être la classification ascendante hiérarchique utilisant la variance minimale comme critère d'agrégation. Les mots doivent être classés par type, la segmentation de l'arbre est effectuée à hauteur constante telle que le nombre de branches soit égal à 1.5 fois le nombre de références souhaitées; les classes les moins importantes sont alors éliminées. Lors de la reconnaissance, c'est la référence la plus proche qui est considérée comme reconnue. On obtient ainsi, sur un vocabulaire difficile (les chiffres français), 89% de bonne reconnaissance en deux propositions avec seulement 2% du volume de données initial.

Mots clés:

Reconnaissance de la parole
Mots isolés
Multilocuteur
Méthode de classification
Classification ascendante hiérarchique