



HAL
open science

Reconnaissance analytique par règles dans SYMULDEPHO, un SYstème MULti-locuteurs de DEcodage acoustico-PHOnétique

Alain Ghio, Mario Rossi

► **To cite this version:**

Alain Ghio, Mario Rossi. Reconnaissance analytique par règles dans SYMULDEPHO, un SYstème MULti-locuteurs de DEcodage acoustico-PHOnétique. Travaux interdisciplinaires du Laboratoire Parole et Langage, 1995, 16, pp.77-92. hal-01665220

HAL Id: hal-01665220

<https://hal.science/hal-01665220v1>

Submitted on 15 Dec 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

RECONNAISSANCE ANALYTIQUE PAR REGLES DANS "SYMULDEPHO", UN *SYSTEME* *MULTI-LOCUTEURS* DE *DECODAGE* ACOUSTICO-*PHONETIQUE*

Résumé

Cette étude s'inscrit dans le cadre de la reconnaissance automatique de la parole et décrit un système multilocuteurs de décodage acoustico-phonétique. Nous souhaitons examiner dans quelle mesure un modèle à base de connaissances phonétiques est capable de décoder de façon automatique la structure phonique de la parole sans recourir aux méthodes stochastiques ou aux informations sémantiques. Le dispositif proposé s'inspire, d'un point de vue fonctionnel, du traitement cognitif humain: le signal de parole alimente non pas un, mais plusieurs modules de décodage fonctionnant en parallèle. Un moteur d'inférence se charge de prendre une décision après une étape de vérification descendante. L'un des modules de reconnaissance analytique est fondé sur l'utilisation de règles phonétiques. Des paramètres acoustiques sont extraits par trames d'analyse de l'ordre de la centiseconde. Les indices et traits phonétiques déduits permettent au système de catégoriser les segments de parole par un ensemble de règles. En outre, un suivi temporel fournit une information contextuelle en analysant les transitions acoustiques de la coarticulation des triphones. Dix locuteurs féminins et masculins ont servi à évaluer le système.

Abstract

The field of this study is automatic speech recognition and concerns more precisely speaker independent acoustic-phonetic decoding. We aim to examine to what extent a knowledge-based model can recognize segmental structure without feedback to semantic information or stochastic modelling. The system proposed is inspired by some features of human cognitive processing: the speech signal activates parallel distributed processes of decoding. A supervisor process takes the final decision after a top-down verification. Analytic recognition is based on phonetic rules. Acoustic cues are extracted at each temporal frame (10 ms). The analysis of phonetic features allows the system to identify the phonemes using rules. A temporal tracking provides contextual information by analyzing acoustic transitions of coarticulation in triphones. Results are given using corpora recorded by 10 male and female french speakers.

RECONNAISSANCE ANALYTIQUE PAR REGLES DANS "SYMULDEPHO", UN **SYSTEME** **MULTI-LOCUTEURS** DE **DECODAGE** ACOUSTICO-**PHONETIQUE**

Alain Ghio et Mario Rossi

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

Introduction

Les technologies vocales peuvent être un excellent banc d'essai pour tester et valider les connaissances que les phonéticiens apportent sur la communication parlée. La reconnaissance automatique de la parole représente, à ce sujet-là, un immense défi. Le cadre de cette étude concerne plus particulièrement le décodage acoustico-phonétique. Notre objectif actuel n'est pas de fournir un système complet et performant à court terme; nous souhaitons examiner dans quelle mesure un modèle à base de connaissances phonétiques est capable de décoder de façon automatique la structure phonique de la parole sans recourir ni aux méthodes stochastiques, ni aux informations sémantiques.

1. Le système de décodage "SYMULDEPHO"

1.1. Présentation générale

SYMULDEPHO (Ghio et Rossi, 1993) est un SYstème MULtilocuteurs de DEcodage acoustico-PHOnétique qui s'inspire, d'un point de vue fonctionnel, du traitement cognitif humain: le signal de parole alimente non pas un, mais plusieurs modules de décodage, qui varient à la fois par le traitement qu'ils opèrent sur le signal et par la conception de l'analyse. Le principe est de faire fonctionner en parallèle les différents modules, chacun d'entre eux analysant le signal de parole de façon indépendante. Dans une approche multi-modale, on pourrait éventuellement associer en juxtaposition avec le canal oral, un décodage du canal visuel. Une telle démarche permet d'exploiter les redondances de la communication parlée, chaque module de décodage fournissant un résultat conforté ou infirmé par ses voisins. De plus, elle autorise la levée d'ambiguïtés et la résolution de problèmes parfois non résolus dans un fonctionnement linéaire.

Un moteur d'inférence se charge de prendre une décision finale après une étape de vérification descendante. Ce type de fonctionnement est déjà bien connu (Stern et al., 1986; Carbonnel et al, 1986 ; Fohr, 1986). Nous nous orientons dans la même direction.

1.2. L'architecture de "SYMULDEPHO"

Les modules d'extraction de paramètres acoustiques fournissent des informations de bas niveau (énergies, taux de passage par zéro, voisement, analyse spectrale...). Ces calculs sont suivis d'une évaluation d'indices et de traits phonétiques qui affinent l'analyse. Les extractions sont diversement utilisées selon les modules de décodage, du fait de leurs différences de conception et de fonctionnement. Les modules de décodage sont les suivants (Figure 1) :

- Un *module de segmentation automatique* fournit, suivant le modèle de Construction de Niveaux (Level Building), un ensemble hiérarchisé de propriétés et de segments acoustiques et phonétiques congruents avec les unités phonétiques et leur structure interne (Rossi, 1990). Le résultat de la segmentation est un étiquetage des trames d'analyse en macro-classes (voyelle, occlusive, constrictive, consonne vocalique, silence...), ce qui permet un repérage des unités et une analyse orientée.
- La *méthode de reconnaissance globale*, qui est de type métrique, s'attache à décoder des segments de type CV (Ghio et Rossi, 1994). Pour cela, dans un premier temps, le module utilise des caractéristiques acoustiques en effectuant une analyse de type P.L.P. (Prédiction Linéaire sur un spectre Perceptif) (Yong & Mason, 1987 ; Hermansky, 1990). Une technique d'alignement dynamique temporel (Data Time Warping ou DTW) permet ensuite de comparer les paramètres acoustiques du stimulus à ceux de prototypes archivés préalablement. Finalement, une étude fine du tableau des prétendants permet de fournir une liste classée de couples CV candidats.
- Un *premier module de reconnaissance analytique* utilise un ensemble de réseaux qui sont des graphes orientés à transitions d'état (Ghio et Rossi, 1994). Ceux-ci sont construits pour modéliser tous les allophones d'une voyelle et non des unités abstraites. Chaque réseau, spécialisé pour la reconnaissance d'une voyelle, est stimulé à chaque trame d'analyse. Le résultat apparaît sous la forme d'une table contenant les allophones candidats. L'analyse de la distribution des candidats conduit à des hypothèses fortes sur l'identification de la voyelle ainsi que de son contexte.

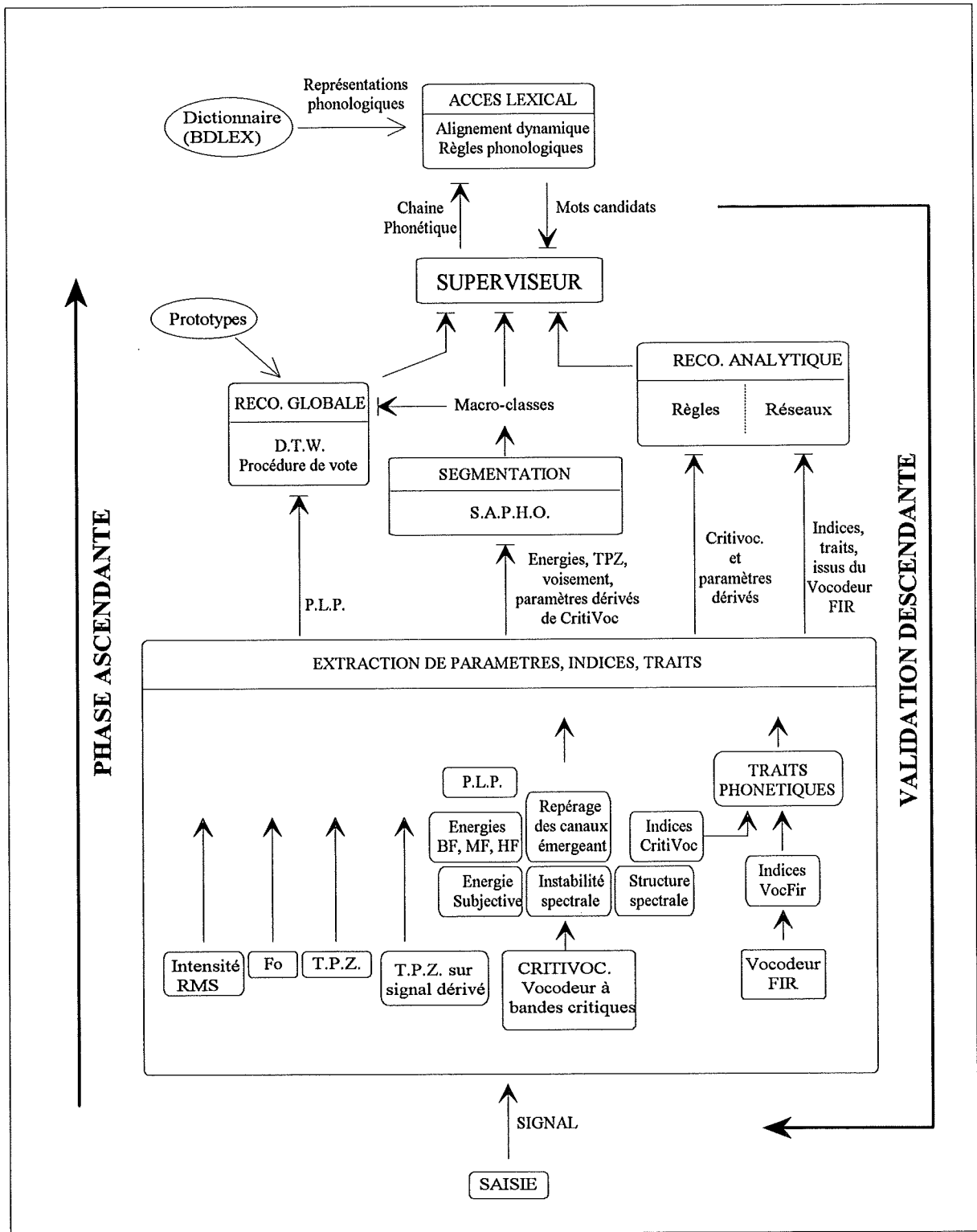


Figure 1: Schéma fonctionnel de "SYMULDEPHO"

- Le *second module de reconnaissance analytique* est fondé sur l'utilisation de règles phonétiques. Des paramètres acoustiques sont extraits par trames d'analyse de l'ordre de la centiseconde. Les indices et traits phonétiques déduits permettent au système de catégoriser, par un ensemble de règles, les segments de parole. Dans le présent document, nous exposerons les premiers résultats.

2. L'extraction de l'information acoustique

'*CritiVoc*' est un vocodeur qui utilise une modélisation de phénomènes psycho-acoustiques (pondération sonore, intégration par bandes critiques) et effectue une représentation compacte temps / fréquence du signal de parole. Dans un premier temps, le traitement auditif effectué par '*CritiVoc*' (Figure 2) consiste à corriger le signal d'un point de vue spectral en appliquant une pondération sonore sur le spectre brut pour tenir compte de la non-linéarité fréquentielle de l'oreille. Dans un second temps, on réalise une intégration du spectre par bandes critiques (Zwicker & Therhart., 1980).

Pour une trame de signal, une telle analyse fournit une série de valeurs correspondant à l'énergie calculée dans chaque bande critique. A une bande passante de 8 kHz correspondent 21 bandes critiques. Renouvelée dans le temps, l'extraction permet d'obtenir une représentation de type "vocodeur", sur laquelle s'appuie la reconnaissance analytique par règles.

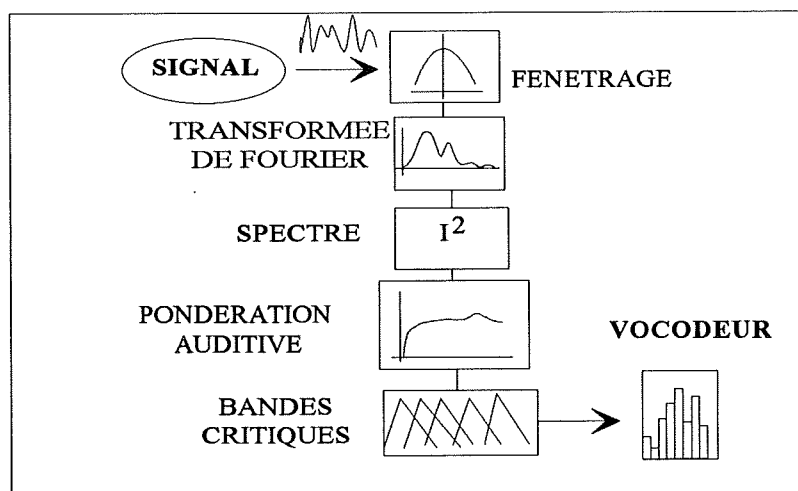


Figure 2 : Extraction de l'information acoustique

3. Le principe de la reconnaissance analytique par règles

3.1. Une première classification des voyelles par traits

L'évaluation des énergies obtenues par le vocodeur à bandes critiques permet la déduction de deux traits fondamentaux: un trait acoustique de gravité/acuité et un trait articuloire d'ouverture/fermeture. Le trait grave/aigu s'obtient à partir d'indices de comparaison entre les énergies de moyennes fréquences par rapport aux hautes fréquences. Le trait d'ouverture s'estime en testant les canaux situés entre 200 Hz et 800 Hz. Cette analyse permet d'effectuer une catégorisation en 4 classes :

- voyelles ouvertes/aigues : /œ/, /ɛ/, /a¹, /ɛ̃/
- voyelles ouvertes/graves : /œ/, /ɔ/, /a², /ã/, /õ/
- voyelles fermées/aigues : /y/, /i/, /e/, /ø/
- voyelles fermées/graves : /u/, /o/

3.2. Une analyse des émergences dans le spectre en bandes critiques

Une fois la catégorisation obtenue, le module effectue une recherche des émergences pertinentes dans la distribution spectrale en bandes critiques. Des problèmes peuvent apparaître pour les voix de femmes particulièrement aigues du fait de l'apparition, dans l'organisation spectrale, d'une forte émergence des harmoniques de la fréquence fondamentale. La détection et la correction d'un tel phénomène est donc nécessaire. Finalement, en tenant compte de la classe vocalique et de l'emplacement des bandes émergentes, un ou plusieurs candidats sont proposés à chaque trame d'analyse. Un filtrage des résultats permet d'éliminer l'apparition isolée d'un candidat (Figure 3).

Dans les résultats présentés ci-dessous, aucune des règles permettant de prendre en compte la variabilité contextuelle n'est active. Aucun suivi dynamique n'est effectué. Seule l'analyse statique trame à trame entre en jeu.

¹en contexte vélaire

²hors contexte vélaire

4. Les résultats

4.1. Les conditions de l'évaluation

Les tests ont porté sur les sept voyelles orales [a,i,u,e,y,o,ø] issues du corpus acoustique SYL de BD-SON (Cervantès et al, 1986). Les énoncés apparaissent sous la forme de logatomes du type CVCV (ex: titi, bubu...). Nous avons utilisé un contexte consonantique multiple (six occlusives, six constrictives, deux nasales, deux liquides). Dix locuteurs (cinq femmes, cinq hommes) ont été choisis, soit un total de 7 voyelles * 2 réalisations * 16 contextes * 10 locuteurs = 2240 énoncés.

4.2. Les résultats bruts

Le tableau I présente les résultats de façon globale. On y distingue :

- le nombre de tests effectués (Total)
- le nombre de cas où la voyelle est correctement reconnue (Correct)
- le taux de reconnaissance (Taux)
- le nombre de candidats total fournis par le système (Ncnd)
- le nombre moyen de candidats par test (Mcnd)
- un coefficient d'efficacité qui donne le pourcentage de réponse correcte par rapport au nombre total de candidats fournis par le système (Effic)

Tableau I: Résultat global de la reconnaissance analytique

Total	Correct	Taux	Ncnd	Mcnd	Effic
2237	2068	92,4 %	5264	2,35	39,3 %

L'analyse succincte des résultats est la suivante : en moyenne, le dispositif de reconnaissance analytique par règles propose 2,35 candidats parmi les 10 voyelles [a, i, u, o, e, y, ø, œ, ɔ, ε] ; dans 92,4 % des cas, la bonne réponse est parmi les candidats. Ces résultats sont à nuancer en tenant compte de chaque voyelle (Tableau II).

Tableau II: Résultats par voyelle de la reconnaissance analytique

Voyelle	Total	Correct	Taux	Ncnd	Mcnd	Effic
a	320	299	93%	688	2,2	43%
i	320	303	95%	648	2	47%
u	320	292	91%	703	2,2	42%
e	319	307	96%	819	2,6	37%
o	318	256	81%	723	2,3	35%
y	320	306	96%	900	2,8	34%
∅	320	305	95%	783	2,4	39%

Le tableau III illustre le classement des voyelles en fonction du taux de reconnaissance, du nombre moyen de candidats et de l'efficacité du décodage relatifs à chacune d'elle.

Tableau III: Classement des voyelles en fonction
(a) du taux de reconnaissance, (b) du nombre moyen de candidats, (c) de l'efficacité

Voyelle	Taux (a)
e	96%
y	96%
∅	95%
i	95%
a	93%
u	91%
o	81%

Voyelle	Mcnd (b)
i	2
a	2,2
u	2,2
o	2,3
∅	2,4
e	2,6
y	2,8

Voyelle	Effic (c)
i	47%
a	43%
u	42%
∅	39%
e	37%
o	35%
y	34%

La difficulté d'identification des voyelles graves peut s'expliquer par la faiblesse de F_2 et F_3 , ce qui entraîne un contraste très faible dans l'analyse en bandes critiques et donc un manque certain d'information. Les bonnes performances dans le décodage de /y/ et /e/ sont à modérer en tenant compte de l'efficacité. Il semble que ces deux entités aient tendance à faire déclencher un lot important de règles, entraînant un nombre plus élevé de candidats et donc l'apparition de la bonne réponse.

4.3. Les influences

4.3.1. Les effets dus au contexte consonantique

Le tableau IV présente les résultats de la reconnaissance des voyelles en fonction du contexte consonantique, qui est du type CVC. Les colonnes "Correct" et "Taux" ne sont donc pas relatives au décodage des consonnes, mais bien des phonèmes [a,i,u,o,e,y,ø].

Tableau IV: Résultats de la reconnaissance analytique des voyelles en fonction du contexte consonantique

Contexte	Total	Correct	Taux	Ncnd	Mcnd	Effic
p	140	129	92,1%	327	2,34	39,4%
t	140	132	94,3%	355	2,54	37,2%
k	140	133	95%	346	2,47	38,4%
b	140	135	96,4%	318	2,27	42,5%
d	140	137	97,9%	333	2,38	41,1%
g	140	134	95,7%	348	2,49	38,5%
f	140	131	93,6%	319	2,28	41,1%
s	140	128	91,4%	315	2,25	40,6%
ʃ	139	127	91,4%	311	2,24	40,8%
v	140	135	96,4%	329	2,35	41%
z	140	130	92,9%	353	2,52	36,8%
ʒ	140	132	94,3%	354	2,53	37,3%
m	140	127	90,7%	295	2,11	43,1%
n	140	127	90,7%	327	2,34	38,8%
l	140	134	95,7%	326	2,33	41,1%
r	138	97	70,3%	308	2,23	31,5%

De façon très nette, le contexte /r/ se distingue par le nombre d'erreurs qui interviennent dans le décodage des voyelles qui lui sont adjacentes. Sa faculté à déformer son entourage est incontestable. La plupart des erreurs proviennent d'un abaissement des maxima spectraux vers les basses fréquences, ce qui est un phénomène connu. Il conviendra d'en tenir compte à l'avenir. Le tableau V illustre le regroupement des contextes consonantiques par macro-classes.

Tableau V: Résultats de la reconnaissance analytique des voyelles
 en fonction des classes du contexte consonantique
 (a) par contexte de macro-classes, (b) dans le contexte obstruant seulement

Contexte	Total	Correct	Taux
occlusif sourd	420	394	93,8%
occlusif sonore	420	406	96,7%
fricatif sourd	419	386	92,1%
fricatif sonore	420	397	94,5%
nasal	280	254	90,7%
l	140	134	95,7%
r	138	97	70,3%

Contexte	Total	Correct	Taux
occlusif	840	800	95,2%
fricatif	839	783	93,3%
sonore	840	803	95,6%
sourd	839	780	93%

La nasalité et l'aspect non voisé du contexte consonantique entraîne des difficultés dans le décodage des voyelles adjacentes.

4.3.2. Les effets dus au locuteur

Le tableau VI présente les résultats de la reconnaissance des voyelles en fonction des locuteurs.

Tableau VI: Résultats de la reconnaissance analytique des voyelles
 en fonction du locuteur

Locuteur	Total	Correct	Taux
lt (f)	224	217	96,9%
bp (m)	224	214	95,5%
nc (f)	222	210	94,6%
rs (f)	224	209	93,3%
jb (m)	224	206	92%
sl (m)	223	204	91,5%
lc (m)	224	204	91,1%
md (f)	224	204	91,1%
po (m)	224	200	89,3%
jo (f)	224	200	89,3%

La variabilité due au locuteur entraîne incontestablement une différence de performance du système. Les résultats médiocres du locuteur "jo" peuvent s'expliquer par le fait que son débit de parole est beaucoup plus rapide que celui des autres locuteurs (20 % au dessus de la moyenne). Les cibles vocaliques sont donc plus difficilement atteintes, ce qui entraîne des erreurs dans l'analyse.

Le tableau VII présente les résultats de la reconnaissance des voyelles en fonction du sexe du locuteur.

Tableau VII: Résultats de la reconnaissance analytique des voyelles en fonction du sexe du locuteur

Sexe	Total	Correct	Taux
f	1118	1040	93 %
m	1119	1028	91,9 %

Il semble se dessiner une tendance légèrement favorable aux voix de femmes, ce qui va à l'encontre de l'idée généralement admise à propos de la difficulté de décodage des voix aiguës.

4.4. Les zones de recouvrement

Dans les tableaux VIIIa et VIIIb sont indiqués le nombre et le pourcentage d'apparitions des voyelles candidates en fonction de la nature du stimulus. Il ne s'agit pas d'une matrice de confusion. Nous faisons référence au phénomène de candidature multiple.

L'information que l'on peut extraire de ce tableau permet d'évaluer les zones de recouvrement des règles. Dans l'ensemble, 5264 candidats (cf.*1) ont été proposés pour 2237 stimuli (cf.*2). Sur 5264 candidats, /a/ est apparu 364 fois (cf.*3). Sur 364 candidatures, 299 (cf.*4) correspondaient à un stimulus /a/, 34 (cf.*5) correspondaient à un stimulus /o/, 23 (cf.*6) correspondaient à un stimulus /ø/...

Tableau VIIIa: Nombre d'apparition des voyelles en tant que candidat en fonction des stimuli présentés

	Stimulus	a	i	u	e	o	y	ø	Total
	Nb	320	320	320	319	318	320	320	2237*2
Candidat									
a		299*4	0	2	6	34*5	0	23*6	364*3
i		1	303	19	148	1	147	17	636
u		0	8	292	1	181	5	27	514
e		17	249	8	307	2	237	86	906
o		10	5	249	0	256	15	118	653
y		5	63	15	205	1	306	141	736
ø		28	9	64	103	91	185	305	785
ɔ		74	0	0	0	74	0	1	149
ɛ		10	0	0	15	0	0	20	45
œ		188	0	0	5	2	0	24	219
xxx		56	11	54	29	81	5	21	257
									5264*1

Le tableau ci-dessous présente les mêmes résultats sous forme de pourcentages. On peut mettre en évidence plusieurs faits :

- les classes de voyelles sont correctement identifiées
 - voyelles aiguës : /i/, /e/, /y/
 - voyelles graves : /a/, /o/
- au sein d'une classe, la confusion entre les voyelles restent non négligeables
 - les règles pour /e/ fonctionnent dans 27% des cas avec /i/ comme stimulus et 26% avec /y/.
 - les règles pour /u/ fonctionnent dans 35% des cas avec /o/ comme stimulus et vice-versa

Tableau VIIIb: Taux d'apparition des voyelles en tant que candidat en fonction des stimuli présentés

	Stimulus	a	i	u	e	o	y	∅	Tot
Candidat									
a		82%	0%	0,5%	1,6%	9,3%	0	6,3%	100%
i		0,2%	48%	3%	23%	0,2%	23%	2,7%	100%
u		0	1,6%	57%	0,2%	35%	1%	5,3%	100%
e		1,9%	27%	0,9%	34%	0,2%	26%	9,5%	100%
o		1,5%	0,8%	38%	0%	39%	2,3%	18%	100%
y		0,7%	8,6%	2%	28%	0,1%	42%	19%	100%
∅		3,6%	1,1%	8,2%	13%	12%	24%	39%	100%
ɔ		50%	0	0	0%	50%	0	0,7%	100%
ɛ		22%	0	0	33%	0	0	44%	100%
œ		86%	0	0	2,3%	0,9%	0	11%	100%
xxx		22%	4,3%	21%	11%	32%	1,9	8,2%	100%

Conclusion

L'analyse montante effectuée par le module de décodage analytique décrit précédemment reste à évaluer en incluant les règles contextuelles qui tiennent compte de la variabilité structurelle. Le cas du contexte /r/ met en évidence les perturbations occasionnées à la voyelle par les unités contiguës. De plus, il faut garder à l'esprit qu'un système qui décode le signal de parole du mot "vocaliser" en une séquence [vokɛlize], ou le mot "casserole" en [kasorol] n'est pas un système qui se trompe. En effet, [ɛ] peut être une forme allophonique de /a/ s'il précède la voyelle /i/, et [o] peut être une variante de /ø/ s'il précède la consonne /r/. Un tel décodage n'est pas incorrect. Un accès lexical et une vérification descendante permettent de lever des ambiguïtés non résolues de façon montante.

Bibliographie

- Carbonnel N., Haton J.P., Fohr D., Lonchamp F., Pierrel J.M. (1986). APHODEX: Design and implementation of an acoustic-phonetic decoding expert system. *Proceedings of IEEE, ICASSP*, Tokyo, Vol.2, 1201-1204.
- Cervantès O., Sérignat J.F., Descout R., Carré R. (1986). "Définition et réalisation d'une base de données des sons du français", *Actes des 15èmes Journées d'Etudes sur la Parole*, GALF, Aix-en-Provence, 213-216
- Fohr D. (1986). *APHODEX: un système expert en décodage acoustico-phonétique de la parole continue*. Thèse de doctorat, Université de Nancy I.

- Ghio A., Rossi M. (1993). SYMULDEPHO : un SYstème MULti-locuteurs de DEcodage acoustico-PHOnétique, *Travaux de l'Institut de Phonétique d'Aix*, Vol.15, 157-184
- Ghio A., Rossi M. (1994). Reconnaissance globale et analytique dans SYMULDEPHO, un SYstème MULti-locuteurs de DEcodage acoustico-PHOnétique, *Actes du séminaire "Reconnaissance automatique de la parole"*, Nancy, GDR-PRC Communication Homme-Machine
- Hermansky H. (1990). Perceptual linear predictive (PLP) analysis of speech . *J.Acoust.Soc.Am.*, 87, 1738-1752.
- Rossi, M. (1990). Segmentation automatique de la parole : Pourquoi ? Quel segments ? *Revue Traitement du signal, GRETSI*, numéro spécial, 315-326.
- Stern, P.E., Eskenazi, M., Memmi D. (1986). An expert system for speech spectrogram reading. *Proceedings of IEEE ICASSP*, Tokyo, 23.1.1- 23.1.4.
- Yong G., Mason J.S. (1987). A comparison between vocal tract and auditory feature analysis in ASR. *Proceedings of Eurospeech*, Edinburgh, 132-135.
- Zwicker E., Therhart E. (1980). Analytical expression for critical bands rate and critical bandwidth as a function of frequency. *J.Acoust.Soc.Am.*, 68, 1523-1525.