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

Alain Ghio, Mario Rossi

RECONNAISSANCE ANALYTIQUE PAR REGLES DANS "SYMULDEPHO", UN *SY*STEME *MUL*TI-LOCUTEURS DE *DE*CODAGE ACOUSTICO-*PHO*NETIQUE

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 *SY*STEME *MUL*TI-LOCUTEURS DE *DE*CODAGE ACOUSTICO-*PHO*NETIQUE

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.

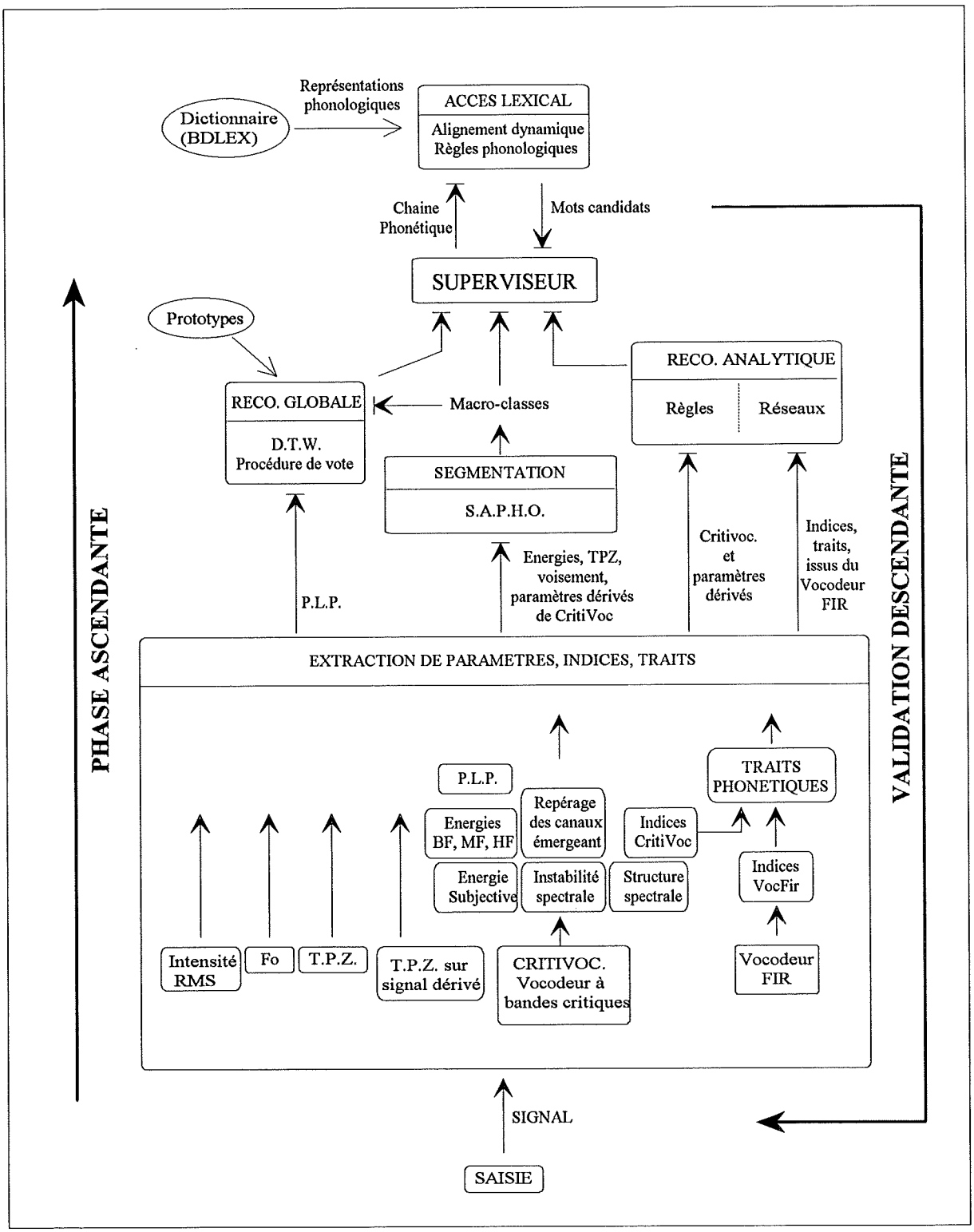


Figure 1: Schéma fonctionnel de "SYMULDEPHO"

• 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 sonique, 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 sonique sur le spectre brut pour tenir compte de la non-linéarité fréquentielle de l'oreille. Dans un second temps, on réalise une intégration du spectre par bandes critiques (Zwicker & Therhart., 1980).

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.

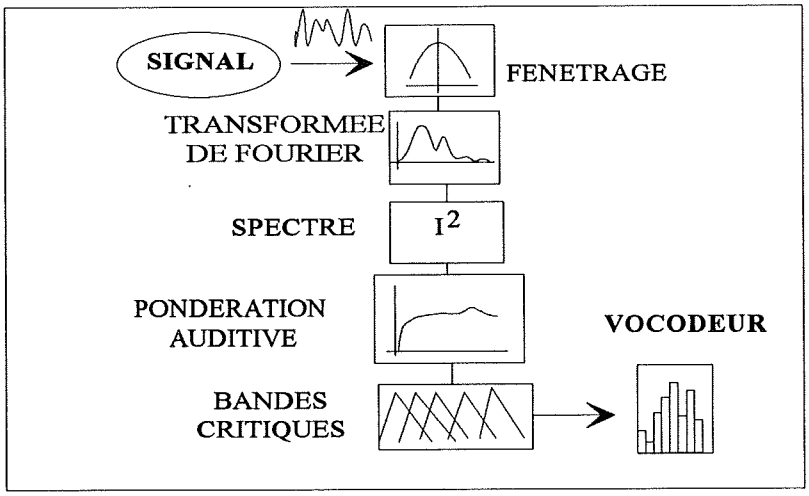


Figure 2 : Extraction de l'information acoustique

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 articulatoire 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/, /a/[1], /ɛ̃/
- voyelles ouvertes/graves : /œ/, /ɔ/, /a/[2], /ɑ̃/, /ɔ̃/
- voyelles fermées/aigues : /y/, /i/, /e/, /ø/
- voyelles fermées/graves : /u/, /o/

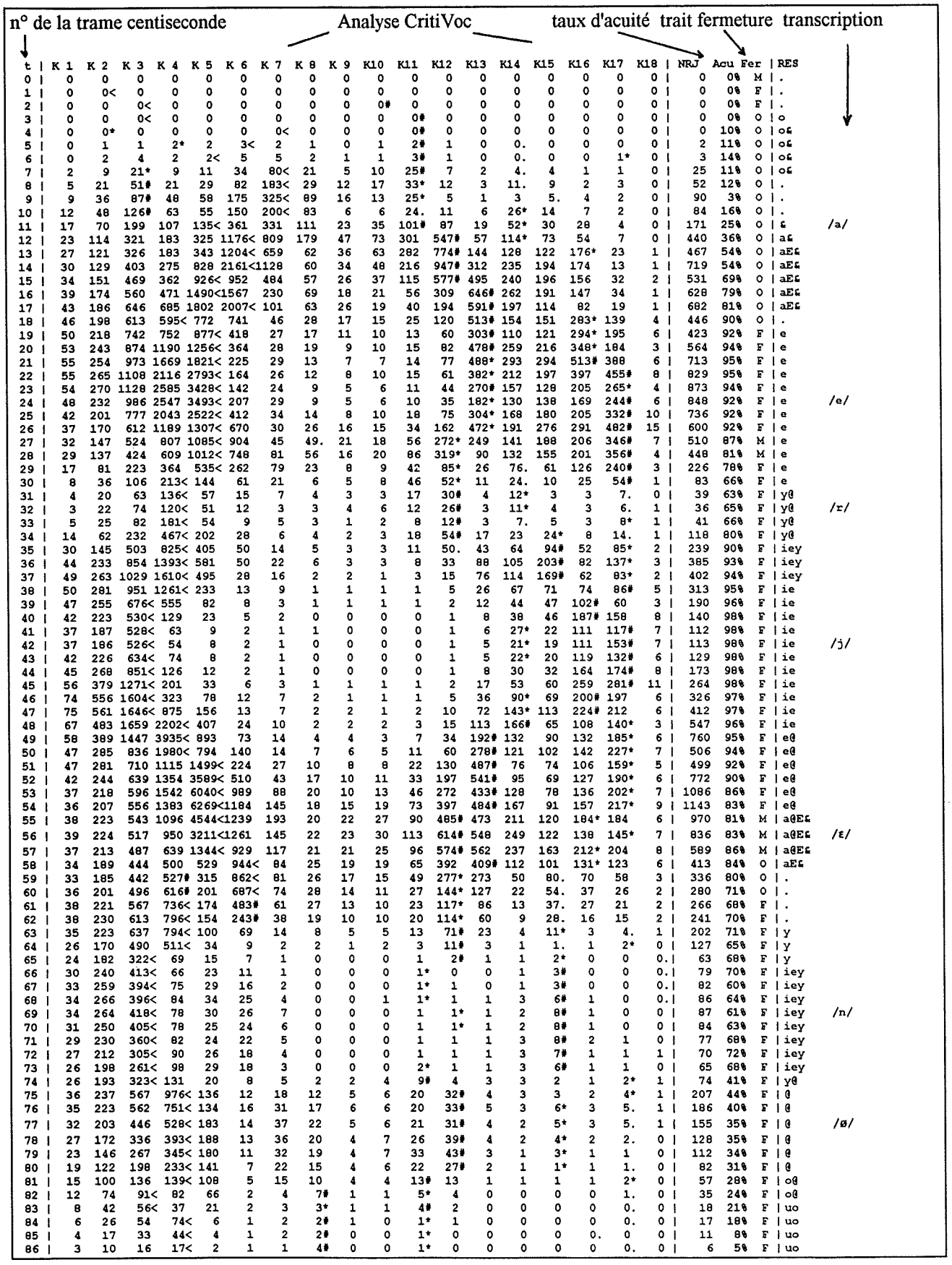
3.2. Une analyse des émergences dans le spectre en bandes critiques

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.

---

[1]en contexte vélaire
[2]hors contexte vélaire



t	K 1	K 2	K 3	K 4	K 5	K 6	K 7	K 8	K 9	K10	K11	K12	K13	K14	K15	K16	K17	K18	NRJ	Acu	Fer	RES	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0%	M	.	
1	0	0<	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0%	F	.	
2	0	0	0<	0	0	0	0	0	0	0#	0	0	0	0	0	0	0	0	0	0%	F	.	
3	0	0	0<	0	0	0	0	0	0	0	0#	0	0	0	0	0	0	0	0	0%	O	o	
4	0	0*	0	0	0	0	0<	0	0	0	0#	0	0	0	0	0	0	0	0	10%	O	o&	
5	0	1	1	2*	2	3<	2	1	0	1	2#	1	0	0.	0	0	0	0	2	11%	O	o&	
6	0	2	4	2	2<	5	5	2	1	1	3#	1	0	0.	0	0	1*	0	3	14%	O	o&	
7	2	9	21*	9	11	34	80<	21	5	10	25#	7	2	4.	4	1	1	0	25	11%	O	o&	
8	5	21	51#	21	29	82	183<	29	12	17	33*	12	3	11.	9	2	3	0	52	12%	O	.	
9	9	36	87#	48	58	175	325<	89	16	13	25*	5	1	3	5.	4	2	0	90	3%	O	.	
10	12	48	126#	63	55	150	200<	83	6	6	24.	11	6	26*	14	7	2	0	84	16%	O	.	
11	17	70	199	107	135<	361	331	111	23	35	101#	87	19	52*	30	28	4	0	171	25%	O	&	/a/
12	23	114	321	183	325	1176<	809	179	47	73	301	547#	57	114*	73	54	7	0	440	36%	O	a&	
13	27	121	326	183	343	1204<	659	62	36	63	282	774#	144	128	122	176*	23	1	467	54%	O	aE&	
14	30	129	403	275	828	2161<	1128	60	34	48	216	947#	312	235	194	174	13	1	719	54%	O	aE&	
15	34	151	469	362	926<	952	484	57	26	37	115	577#	495	240	196	156	32	2	531	69%	O	aE&	
16	39	174	560	471	1490<	1567	230	69	18	21	56	309	646#	262	191	147	34	1	628	79%	O	aE&	
17	43	186	646	685	1802	2007<	101	63	26	19	40	194	591#	197	114	82	19	1	682	81%	O	aE&	
18	46	198	613	595<	772	741	46	28	17	15	25	120	513#	154	151	283*	139	4	446	90%	O	.	
19	50	218	742	752	877<	418	27	17	11	10	13	60	303#	110	121	294*	195	6	423	92%	F	e	
20	53	243	874	1190	1256<	364	28	19	9	10	15	82	478#	259	216	348*	184	3	564	94%	F	e	
21	55	254	973	1669	1821<	225	29	13	7	7	14	77	488*	293	294	513#	388	6	713	95%	F	e	
22	55	265	1108	2116	2793<	164	26	12	8	10	15	61	382*	212	197	397	455#	8	829	95%	F	e	
23	54	270	1128	2585	3428<	142	24	9	5	6	11	44	270#	157	128	205	265*	4	873	94%	F	e	
24	48	232	986	2547	3493<	207	29	9	5	6	10	35	182*	130	138	169	244#	6	848	92%	F	e	/e/
25	42	201	777	2043	2522<	412	34	14	8	10	18	75	304*	168	180	205	332#	10	736	92%	F	e	
26	37	170	612	1189	1307<	670	30	26	16	15	34	162	472*	191	276	291	482#	15	600	92%	F	e	
27	32	147	524	807	1085<	904	45	49.	21	18	56	272*	249	141	188	206	346#	7	510	87%	M	e	
28	29	137	424	609	1012<	748	81	56	16	20	86	319*	90	132	155	201	356#	4	448	81%	M	e	
29	17	81	223	364	535<	262	79	23	8	9	42	85*	26	76.	61	126	240#	3	226	78%	F	e	
30	8	36	106	213<	144	61	21	6	5	8	46	52*	11	24.	10	25	54#	1	83	66%	F	e	
31	4	20	63	136<	57	15	7	4	3	3	17	30#	4	12*	3	3	7.	0	39	63%	F	y@	
32	3	22	74	120<	51	12	3	3	4	6	12	26#	3	11*	4	3	6.	1	36	65%	F	y@	/ɛ/
33	5	25	82	181<	54	9	5	3	1	2	8	12#	3	7.	5	3	8*	1	41	66%	F	y@	
34	14	62	232	467<	202	28	6	4	2	3	18	54#	17	23	24*	8	14.	1	118	80%	F	y@	
35	30	145	503	825<	405	50	14	5	3	3	11	50.	43	64	94#	52	85*	2	239	90%	F	iey	
36	44	233	854	1393<	581	50	22	6	3	3	8	33	88	105	203#	82	137*	3	385	93%	F	iey	
37	49	263	1029	1610<	495	28	16	2	2	1	3	15	76	114	169#	62	83*	2	402	94%	F	iey	
38	50	281	951	1261<	233	13	9	1	1	1	1	5	26	67	71	74	86#	5	313	95%	F	ie	
39	47	255	676<	555	82	8	3	1	1	1	1	2	12	44	47	102#	60	3	190	96%	F	ie	
40	42	223	530<	129	23	5	2	0	0	0	0	1	8	38	46	187#	158	8	140	98%	F	ie	
41	37	187	528<	63	9	2	1	1	0	0	0	1	6	27*	22	111	117#	7	112	98%	F	ie	
42	37	186	526<	54	8	2	1	0	0	0	0	1	5	21*	19	111	153#	7	113	98%	F	ie	/j/
43	42	226	634<	74	8	2	1	0	0	0	0	1	5	22*	20	119	132#	6	129	98%	F	ie	
44	45	268	851<	126	12	2	1	0	0	0	0	1	8	30	32	164	174#	8	173	98%	F	ie	
45	56	379	1271<	201	33	6	3	1	1	1	1	2	17	53	60	259	281#	11	264	98%	F	ie	
46	74	556	1604<	323	78	12	7	2	1	1	1	5	36	90*	69	200#	197	6	326	97%	F	ie	
47	75	561	1646<	875	156	13	7	2	2	1	2	10	72	143*	113	224#	212	6	412	97%	F	ie	
48	67	483	1659	2202<	407	24	10	2	2	2	3	15	113	166#	65	108	140*	3	547	96%	F	ie	
49	58	389	1447	3935<	893	73	14	4	4	3	7	34	192#	132	90	132	185*	6	760	95%	F	e@	
50	47	285	836	1980<	794	140	14	7	6	5	11	60	278#	121	102	142	227*	7	506	94%	F	e@	
51	47	281	710	1115	1499<	224	27	10	8	8	22	130	487#	76	74	106	159*	5	499	92%	F	e@	
52	42	244	639	1354	3589<	510	43	17	10	11	33	197	541#	95	69	127	190*	6	772	90%	F	e@	
53	37	218	596	1542	6040<	989	88	20	10	13	46	272	433#	128	78	136	202*	7	1086	86%	F	e@	
54	36	207	556	1383	6269<	1184	145	18	15	19	73	397	484#	167	91	157	217*	9	1143	83%	F	e@	
55	38	223	543	1096	4544<	1239	193	20	22	27	90	485#	473	211	120	184*	184	6	970	81%	M	a@E&	
56	39	224	517	950	3211<	1261	145	22	23	30	113	614#	548	249	122	138	145*	7	836	83%	M	a@E&	/ɛ/
57	37	213	487	639	1344<	929	117	21	21	25	96	574#	562	237	163	212*	204	8	589	86%	M	a@E&	
58	34	189	444	500	529	944<	84	25	19	19	65	392	409#	112	101	131*	123	6	413	84%	O	aE&	
59	33	185	442	527#	315	862<	81	26	17	15	49	277*	273	50	80.	70	58	3	336	80%	O	.	
60	36	201	496	616#	201	687<	74	28	14	11	27	144*	127	22	54.	37	26	2	280	71%	O	.	
61	38	221	567	736<	174	483#	61	27	13	10	23	117*	86	13	37.	27	21	2	266	68%	F	.	
62	38	230	613	796<	154	243#	38	19	10	10	20	114*	60	9	28.	16	15	2	241	70%	F	.	
63	35	223	637	794<	100	69	14	8	5	5	13	71#	23	4	11*	3	4.	1	202	71%	F	y	
64	26	170	490	511<	34	9	2	2	1	2	3	11#	3	1	1.	1	2*	0	127	65%	F	y	
65	24	182	322<	69	15	7	1	0	0	0	1	2#	1	1	2*	0	0	0.	63	68%	F	y	
66	30	240	413<	66	23	11	1	0	0	0	1*	0	0	1	3#	0	0	0.	79	70%	F	iey	
67	33	259	394<	75	29	16	2	0	0	0	1*	1	0	1	3#	0	0	0.	82	60%	F	iey	
68	34	266	396<	84	34	25	4	0	0	1	1*	1	1	3	6#	1	0	0.	86	64%	F	iey	
69	34	264	418<	78	30	26	7	0	0	0	1	1*	1	2	8#	1	0	0	87	61%	F	iey	/n/
70	31	250	405<	78	25	24	6	0	0	0	1	1*	1	2	8#	1	0	0	84	63%	F	iey	
71	29	230	360<	82	24	22	5	0	0	0	1	1	1	3	8#	2	1	0	77	68%	F	iey	
72	27	212	305<	90	26	18	4	0	0	0	1	1	1	3	7#	1	1	1	70	72%	F	iey	
73	26	198	261<	98	29	18	3	0	0	0	2*	1	1	3	6#	1	1	0	65	68%	F	iey	
74	26	193	323<	131	20	8	5	2	2	4	9#	4	3	3	2	1	2*	1	74	41%	F	y@	
75	36	237	567	976<	136	12	18	12	5	6	20	32#	4	3	3	2	4*	1	207	44%	F	@	
76	35	223	562	751<	134	16	31	17	6	6	20	33#	5	3	6*	3	5.	1	186	40%	F	@	
77	32	203	446	528<	183	14	37	22	5	6	21	31#	4	2	5*	3	5.	1	155	35%	F	@	/ø/
78	27	172	336	393<	188	13	36	20	4	7	26	39#	4	2	4*	2	2.	0	128	35%	F	@	
79	23	146	267	345<	180	11	32	19	4	7	33	43#	3	1	3*	1	1	0	112	34%	F	@	
80	19	122	198	233<	141	7	22	15	4	6	22	27#	2	1	1*	1	1.	0	82	31%	F	@	
81	15	100	136	139<	108	5	15	10	4	4	13#	13	1	1	1	1	2*	0	57	28%	F	o@	
82	12	74	91<	82	66	2	4	7#	1	1	5*	4	0	0	0	0	1.	0	35	24%	F	o@	
83	8	42	56<	37	21	2	3	3*	1	1	4#	2	0	0	0	0	0.	0	18	21%	F	uo	
84	6	26	54	74<	6	1	2	2#	1	0	1*	1	0	0	0	0	0.	0	17	18%	F	uo	
85	4	17	33	44<	4	1	2	2#	0	0	1*	0	0	0	0	0.	0	0	11	8%	F	uo	
86	3	10	16	17<	2	1	1	4#	0	0	1*	0	0	0	0	0	0.	0	6	5%	F	uo	

Figure 3 : Répartition spectrale sur le mot "aérienne", locuteur masc.
La colonne RES propose les resultats (E=/ɛ/, @=/ø/, &=/œ/, o= /o/ ou /ɔ/)

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 succinte 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, ɛ] ; 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%
ø	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%
ø	95%
i	95%
a	93%
u	91%
o	81%

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

Voyelle	Effic (c)
i	47%
a	43%
u	42%
ø	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 aigues.

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 fontion 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 aigues : /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 fontion des stimuli présentés

	Stimulus	a	i	u	e	o	y	ø	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%
ø		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 beetween 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.