Chapitre 7

Évaluation du système

Ce chapitre présente les résultats obtenus lors de l'évaluation des différents modules de PLSPP. Nous avons souhaité tester les performances de chaque module de manière isolée, afin de minimiser l'introduction d'erreurs résultant de traitements effectués en amont. Chaque étape d'évaluation a nécessité l'utilisation de données annotées manuellement, permettant une comparaison entre les annotations automatiques produites et une référence.

Pour évaluer la qualité de la segmentation en locuteurs et de la reconnaissance de la parole, nous avons exploité une portion du corpus CLES-FR, dont la transcription et la segmentation en locuteurs ont été vérifiées et corrigées manuellement (*cf.* chapitre 4.6). Nous nous référerons à ce corpus sous le nom de *CLES-gold*. En revanche, l'évaluation de la précision de l'alignement mot-signal a requis un corpus de référence différent, car l'alignement des mots dans le corpus CLES-gold n'a pas été vérifié. À cette fin, nous avons utilisé un corpus de parole spontanée de locuteurs francophones, issu de l'étude de Frost et al. (2024).

L'évaluation de la détection et de l'étiquetage des pauses s'est révélée plus difficile, car les corpus disposant d'une annotation syntaxique des pauses en parole spontanée sont peu nombreux. Nous avons utilisé ici un corpus de dialogues spontanés entre apprenants slovaques de l'anglais, gracieusement fourni par l'université de Nitra en Slovaquie.

Enfin, l'analyse des performances d'annotation de l'accent lexical a été réalisée en la confrontant plusieurs références : la perception d'auditeurs natifs, la conscience phonologique de locuteurs natifs et non natifs, ainsi que les résultats obtenus sur des données de parole native contrôlée (lecture de phrases et de textes).

7.1 Modules de prétraitements

1.1 Segmentation en locuteurs

Pour évaluer la précision de la diarisation en locuteurs obtenue grâce au premier module de PLSPP, nous avons d'abord calculé le taux d'erreur de diarisation (DER) par binôme sur les fichiers de sortie de Pyannote après avoir analysé les 20 discussions du corpus CLES-gold.

Sur les 20 enregistrements du corpus CLES-gold, totalisant 2 h 58 min 38 s de parole, le DER moyen obtenu est de 19,42 % (médiane : 13,21 %, *cf.* figure et tableau 7.1). Ce DER moyen correspond à peu près au taux d'erreur de 18,9 % obtenu par Pyannote sur le corpus d'interactions en réunions professionnelles AMI-IHM *(Augmented Multi-Party Interaction - Individual Headset Microphone)* (Bredin, 2023).

On remarque que deux discussions obtiennent un DER particulièrement élevé (69 % et 64 %), contrastant avec les 18 autres dont la moyenne est de 14,23 % (médiane 10,78 %, min 4 %, max 30 %). On peut constater que le fichier qui obtient un DER de 69 % présente une proportion importante de parole non détectée (59,3 % du temps d'enregistrement). Il semblerait que la voix des locuteurs n'ait pas été correctement détectée par Pyannote pour une raison que nous ne sommes pas parvenus à identifier. Dans le cas du fichier dont le DER est 64 %, il s'agit cette fois d'une importante confusion entre les locuteurs. Comme indiqué dans le tableau 7.1, 39,9 % du temps d'enregistrement de ce fichier n'est pas annoté avec le bon locuteur, alors que la moyenne de durée de confusion sur les autres fichiers n'est que de 3,1 %. Malgré cela, sur l'ensemble des fichiers, on constate que la durée de parole qui n'est pas attribuée au bon locuteur reste limitée : 5,1 % sur la durée totale d'enregistrement, et 4,9 % en moyenne par enregistrement.

Nous avons ensuite calculé l'indice d'interférence I_L de chaque locuteur à partir des segments de parole extraits par PLSPP (durée supérieure ou égale à 8 s). Cet indice donne la proportion de temps de parole d'un locuteur L correspondant en réalité à la parole de l'interlocuteur. C'est un moyen de quantifier la présence de l'interlocuteur dans les segments de parole attribués au locuteur, et qui sera donc potentiellement à l'origine de mauvaises attributions de patterns de pauses ou d'accents lexicaux.

L'indice d'interférence moyen obtenu est de 2,98 % (médiane 1,57 %, min 0 %, max 29,26 %, *cf.* figure 7.2 et tableau 7.2). On retrouve les deux locuteurs de la discussion présentant un pourcentage de confusion important lors du calcul du DER sur la sortie de Pyannote. Ces deux locuteurs ont un indice d'interférence de respectivement 29,26 % et 15,36 %. À l'écoute de l'enregistrement, on constate effectivement de nombreux chevauchements entre les deux locuteurs, à l'origine de ces confusions

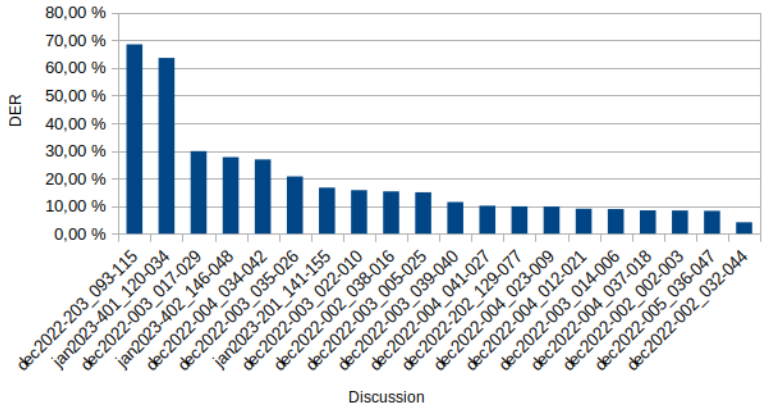


FIG. 7.1 : Taux d'erreurs de diarisation (DER) sur le corpus CLES-gold

	File	DER	Missed Speech		False Alarm		Confusion		Total Duration
1	dec2022-203_093-115	68,54 %	330,98	(59,3 %)	24,31	(4,4 %)	24,33	(4,4 %)	00 :09 :18
2	jan2023-401_120-034	63,62 %	118,62	(19,1 %)	87,94	(14,2 %)	247,92	(39,9 %)	00 :10 :21
3	dec2022-003_017-029	29,87 %	35,15	(8,3 %)	45,92	(10,8 %)	42,69	(10,0 %)	00 :07 :05
4	jan2023-402_146-048	27,70 %	130,13	(22,4 %)	17,02	(2,9 %)	8,7	(1,5 %)	00 :09 :40
5	dec2022-004_034-042	26,85 %	47,69	(9,5 %)	50,6	(10,1 %)	29,32	(5,8 %)	00 :08 :22
6	dec2022-003_035-026	20,72 %	61,94	(10,1 %)	33,41	(5,4 %)	31,33	(5,1 %)	00 :10 :15
7	jan2023-201_141-155	16,66 %	77,73	(13,8 %)	8,71	(1,5 %)	7,36	(1,3 %)	00 :09 :25
8	dec2022-003_022-010	15,75 %	26,3	(8,4 %)	12,04	(3,9 %)	10,03	(3,2 %)	00 :05 :12
9	dec2022-002_038-016	15,30 %	37,91	(6,9 %)	22,81	(4,2 %)	22,72	(4,1 %)	00 :09 :08
10	dec2022-003_005-025	14,97 %	32,21	(6,0 %)	23,83	(4,4 %)	23,83	(4,4 %)	00 :08 :56
11	dec2022-003_039-040	11,45 %	33,68	(4,5 %)	26,27	(3,5 %)	25,44	(3,4 %)	00 :12 :30
12	dec2022-004_041-027	10,11 %	34,29	(8,0 %)	4,83	(1,1 %)	3,65	(0,9 %)	00 :07 :06
13	dec2022-202_129-077	9,87 %	39,25	(7,5 %)	7,14	(1,4 %)	4,14	(0,8 %)	00 :08 :45
14	dec2022-004_023-009	9,80 %	24,41	(4,6 %)	13,56	(2,6 %)	13,56	(2,6 %)	00 :08 :46
15	dec2022-004_012-021	9,03 %	24,94	(4,9 %)	10,88	(2,1 %)	9,98	(2,0 %)	00 :08 :29
16	dec2022-003_014-006	8,91 %	26,22	(4,2 %)	15,28	(2,5 %)	13,28	(2,2 %)	00 :10 :17
17	dec2022-004_037-018	8,43 %	17,31	(4,0 %)	9,91	(2,3 %)	7,91	(1,8 %)	00 :07 :12
18	dec2022-002_002-003	8,38 %	18,93	(3,4 %)	15,65	(2,8 %)	10,51	(1,9 %)	00 :09 :11
19	dec2022-005_036-047	8,26 %	11,94	(2,2 %)	20,84	(3,8 %)	11,12	(2,0 %)	00 :09 :03
20	dec2022-002_032-044	4,09 %	17,56	(3,0 %)	2,94	(0,5 %)	2,94	(0,5 %)	00 :09 :36
			1147,19	**(10,7 %)**	**453,89**	**(4,2 %)**	**550,76**	**(5,1 %)**	**02 :58 :38**

TAB. 7.1 : Taux d'erreurs de diarisation (DER) par discussion sur le corpus CLES-gold. Les durées de parole manquée (Missed Detection), fausse alerte et confusion de locuteur sont données en secondes. Les pourcentages sont calculés par rapport à la durée totale de chaque enregistrement.

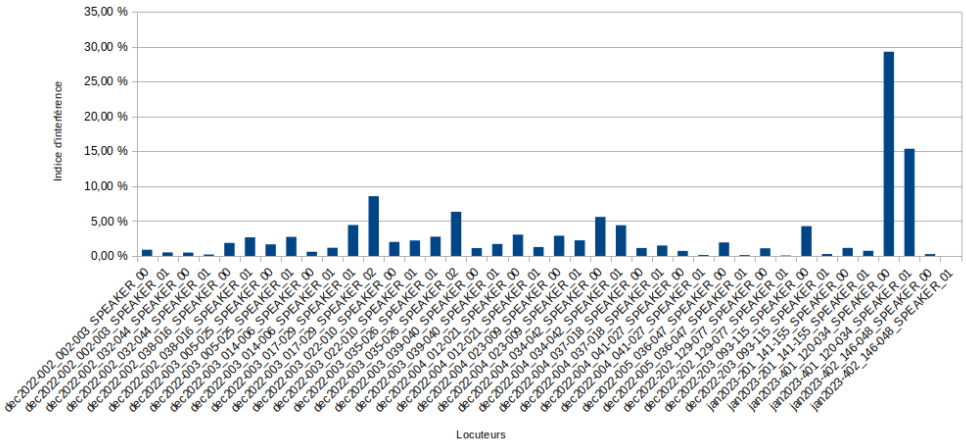


FIG. 7.2 : Indice d'interférence par locuteur (segments de durée supérieure ou égale à 8 s)

d'étiquetage. En revanche, pour la discussion qui a obtenu le DER le plus élevé, l'indice d'interférence pour les deux locuteurs reste assez bas (4,27 % et 0,27 %), mais on voit que la durée de parole extraite est réduite (73 s et 64 s, contre 235 s en moyenne pour les autres locuteurs).

Nous considérons que la proportion d'erreur d'identification du locuteur dans les segments de paroles extraits par PLSPP est suffisamment basse pour ne pas affecter sensiblement les résultats de nos analyses.

1.2 Reconnaissance automatique de la parole

Pour évaluer la qualité de la reconnaissance automatique de la parole avec le système Whisper et le modèle *base.en*[1], nous avons calculé le taux d'erreur de mots (WER) pour chaque segment du corpus CLES-gold (n = 349).

Le WER est la métrique de référence pour évaluer les performances des systèmes de reconnaissance de la parole. Elle permet une comparaison directe entre différents modèles ou systèmes analysant un même corpus. Toutefois, il est essentiel de noter que la précision de la reconnaissance est fortement dépendante du type de parole analysée (Evain, 2024). Par exemple, sur le corpus de textes lus LibriSpeech (environ 1000 h d'enregistrements de livres audio en anglais, Panayotov et al., 2015), le modèle *base.en* obtient un taux d'erreur très faible (4,2 %). En revanche, sur le corpus de réunions professionnelles AMI-IHM cité précédemment, le taux d'erreur s'élève à 20,5 % (Radford et al., 2022, p. 22 pour les WER obtenus par différents modèles de

[1] https://huggingface.co/openai/whisper-base.en

	Speaker	I_L	Duration (s)
1	dec2022-002_002-003_SPEAKER_00	0,88 %	277
2	dec2022-002_002-003_SPEAKER_01	0,50 %	228
3	dec2022-002_032-044_SPEAKER_00	0,48 %	294
4	dec2022-002_032-044_SPEAKER_01	0,20 %	256
5	dec2022-002_038-016_SPEAKER_00	1,86 %	288
6	dec2022-002_038-016_SPEAKER_01	2,66 %	201
7	dec2022-003_005-025_SPEAKER_00	1,65 %	223
8	dec2022-003_005-025_SPEAKER_01	2,72 %	271
9	dec2022-003_014-006_SPEAKER_00	0,59 %	276
10	dec2022-003_014-006_SPEAKER_01	1,16 %	314
11	dec2022-003_017-029_SPEAKER_01	4,43 %	191
12	dec2022-003_017-029_SPEAKER_02	8,56 %	176
13	dec2022-003_022-010_SPEAKER_00	2,00 %	165
14	dec2022-003_022-010_SPEAKER_01	2,21 %	100
15	dec2022-003_035-026_SPEAKER_01	2,75 %	395
16	dec2022-003_035-026_SPEAKER_02	6,33 %	106
17	dec2022-003_039-040_SPEAKER_00	1,12 %	306
18	dec2022-003_039-040_SPEAKER_01	1,71 %	394
19	dec2022-004_012-021_SPEAKER_00	3,05 %	209
20	dec2022-004_012-021_SPEAKER_01	1,27 %	279
21	dec2022-004_023-009_SPEAKER_00	2,89 %	215
22	dec2022-004_023-009_SPEAKER_01	2,23 %	269
23	dec2022-004_034-042_SPEAKER_00	5,59 %	235
24	dec2022-004_034-042_SPEAKER_01	4,40 %	217
25	dec2022-004_037-018_SPEAKER_00	1,13 %	128
26	dec2022-004_037-018_SPEAKER_01	1,49 %	266
27	dec2022-004_041-027_SPEAKER_00	0,71 %	228
28	dec2022-004_041-027_SPEAKER_01	0,13 %	129
29	dec2022-005_036-047_SPEAKER_00	1,93 %	280
30	dec2022-005_036-047_SPEAKER_01	0,12 %	225
31	dec2022-202_129-077_SPEAKER_00	1,09 %	266
32	dec2022-202_129-077_SPEAKER_01	0,06 %	176
33	dec2022-203_093-115_SPEAKER_00	4,27 %	73
34	dec2022-203_093-115_SPEAKER_01	0,27 %	64
35	jan2023-201_141-155_SPEAKER_00	1,14 %	223
36	jan2023-201_141-155_SPEAKER_01	0,73 %	205
37	jan2023-401_120-034_SPEAKER_00	29,26 %	232
38	jan2023-401_120-034_SPEAKER_01	15,36 %	345
39	jan2023-402_146-048_SPEAKER_00	0,26 %	268
40	jan2023-402_146-048_SPEAKER_01	0,00 %	83

Tab. 7.2 : *Indice d'interférence par locuteur et durée totale de parole (segments de durée supérieure ou égale à* 8 s*)*

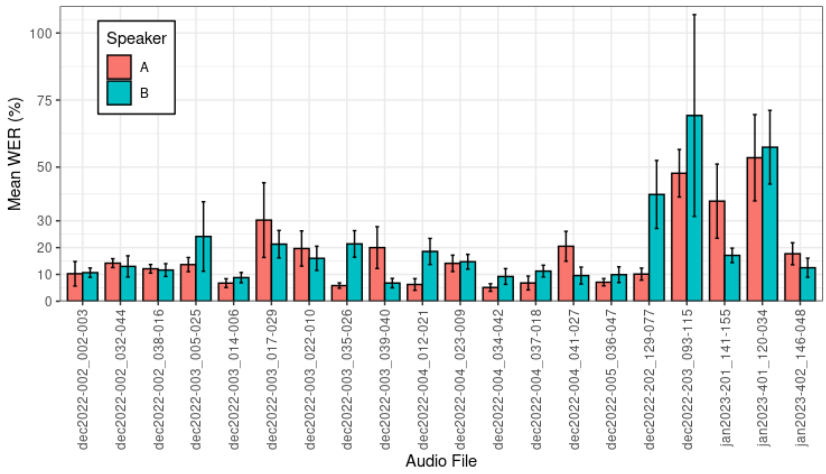


FIG. 7.3 : Taux d'erreur de mots (WER) moyen par locuteur avec barre d'erreur indiquant le degré de variabilité selon les segments du même locuteur (écart type)

Whisper sur divers corpus). Ainsi, la précision du modèle dépend de la situation de parole, et notamment du degré de spontanéité et de la présence de chevauchements.

Pour notre corpus CLES-gold, le WER moyen de l'ensemble des 349 segments est de 19,0 % (médiane : 11,0 %). Le WER par locuteur, calculé par concaténation des segments, est de 16,76 % (médiane : 13,5 %) avec une amplitude allant de 6 % à 63 %. La figure 7.3 et le tableau 7.3 indiquent le WER moyen par locuteur, ainsi que le nombre de substitutions, de délétions et d'insertions.

Les quatre locuteurs qui obtiennent les WER les plus élevés sont ceux des deux enregistrements au DER élevé, identifiés dans la section précédente (*dec2022-203_093-115* et *jan2023-401_120-034*). On constate notamment que le locuteur *jan2023-401_120-034_SPEAKER_00* a un très grand nombre d'insertions (135, contre 16 en moyenne pour les autres). En observant de plus près les segments de ce locuteur, on trouve des phénomènes d'hallucination de Whisper, dus à des répétitions successives de mots. Par exemple, pour le segment *jan2023-401_120-034_SPEAKER_00_2* (57 mots, WER : 186 %, $S = 51$, $D = 0$, $I = 55$), la transcription de référence commence par "*you can you can you can write wait wait wait wait wait we have to write an article* [...]", mais la transcription automatique ne contient que le mot "*wait*" répété 112 fois. Dans d'autres cas, il s'agit d'une interférence importante de l'interlocuteur, aboutissant à un grand nombre d'insertions et de substitutions. Pour les deux locuteurs de la discussion *dec2022-203_093-115*, il s'agit principalement d'une mauvaise reconnaissance générale de la parole du SPEAKER_00, combiné à la quantité de parole limitée pour ce binôme.

	Speaker	WER	SR	DR	IR	Nb of Words
1	dec2022-002_002-003_SPEAKER_00	11	4,55	2,48	4,14	483
2	dec2022-002_002-003_SPEAKER_01	10	4,74	2,91	2,73	549
3	dec2022-002_032-044_SPEAKER_00	15	6,91	2,56	5,37	391
4	dec2022-002_032-044_SPEAKER_01	11	5,51	2,90	2,90	345
5	dec2022-002_038-016_SPEAKER_00	11	6,27	1,76	3,33	510
6	dec2022-002_038-016_SPEAKER_01	10	3,56	4,35	2,37	253
7	dec2022-003_005-025_SPEAKER_00	13	3,74	6,98	2,49	401
8	dec2022-003_005-025_SPEAKER_01	19	3,32	4,34	11,48	392
9	dec2022-003_014-006_SPEAKER_00	6	1,35	2,03	2,36	592
10	dec2022-003_014-006_SPEAKER_01	8	3,21	2,92	2,34	685
11	dec2022-003_017-029_SPEAKER_01	19	5,84	3,11	9,73	257
12	dec2022-003_017-029_SPEAKER_02	20	7,10	2,78	9,88	324
13	dec2022-003_022-010_SPEAKER_00	20	5,15	12,02	3,00	233
14	dec2022-003_022-010_SPEAKER_01	15	7,98	3,07	3,68	163
15	dec2022-003_035-026_SPEAKER_01	6	4,10	1,28	0,77	781
16	dec2022-003_035-026_SPEAKER_02	21	6,94	5,56	8,33	216
17	dec2022-003_039-040_SPEAKER_00	14	3,96	3,79	6,54	581
18	dec2022-003_039-040_SPEAKER_01	6	1,24	3,19	1,44	971
19	dec2022-004_012-021_SPEAKER_00	7	2,56	2,96	0,99	507
20	dec2022-004_012-021_SPEAKER_01	15	4,97	5,33	4,97	563
21	dec2022-004_023-009_SPEAKER_00	15	7,50	3,61	3,61	360
22	dec2022-004_023-009_SPEAKER_01	15	5,86	3,70	4,94	324
23	dec2022-004_034-042_SPEAKER_00	6	1,81	3,62	0,36	276
24	dec2022-004_034-042_SPEAKER_01	11	4,10	4,10	2,52	317
25	dec2022-004_037-018_SPEAKER_00	6	2,15	1,72	2,58	233
26	dec2022-004_037-018_SPEAKER_01	11	5,37	3,74	2,10	428
27	dec2022-004_041-027_SPEAKER_00	20	10,29	3,22	6,75	311
28	dec2022-004_041-027_SPEAKER_01	7	3,47	0,99	2,97	202
29	dec2022-005_036-047_SPEAKER_00	7	4,42	1,52	1,37	656
30	dec2022-005_036-047_SPEAKER_01	9	4,99	2,49	1,59	441
31	dec2022-202_129-077_SPEAKER_00	11	6,38	2,26	2,67	486
32	dec2022-202_129-077_SPEAKER_01	27	8,53	5,81	12,40	258
33	dec2022-203_093-115_SPEAKER_00	45	29,59	8,16	7,14	98
34	dec2022-203_093-115_SPEAKER_01	45	9,78	7,61	27,17	92
35	jan2023-201_141-155_SPEAKER_00	28	12,61	6,08	9,23	444
36	jan2023-201_141-155_SPEAKER_01	17	9,46	5,68	1,62	370
37	jan2023-401_120-034_SPEAKER_00	63	22,85	3,76	36,29	372
38	jan2023-401_120-034_SPEAKER_01	45	26,32	8,30	10,32	494
39	jan2023-402_146-048_SPEAKER_00	16	7,22	5,35	3,48	374
40	jan2023-402_146-048_SPEAKER_01	13	9,62	1,92	1,92	156
	Mean :	**16,85**	**7,13**	**4,00**	**5,75**	

Tab. 7.3 : *Taux d'erreur de mots (WER), de substitutions (SR), de délétions (DR), et d'insertions (IR), et nombre total de mots par locuteur*

En excluant ces quatre locuteurs, le WER moyen des 36 locuteurs restants est de 12,91 % (médiane : 11,0 %, min : 6 %, max : 28 %). Ces observations mettent en évidence une certaine variabilité du WER selon les locuteurs, mais les résultats restent globalement satisfaisants compte tenu de la nature de la parole analysée, spontanée et L2. Ces taux d'erreurs pourraient toutefois être réduits en utilisant un modèle de langue avec plus de paramètres, comme *medium.en*[2] (769 millions de paramètres contre 74 millions pour *base.en*).

1.3 Alignement mot-signal

Pour évaluer la précision de l'alignement temporel des mots de la transcription orthographique au signal de parole, nous avons mesuré les performances de l'aligneur utilisé par PLSPP, Wav2Vec2.0, en nous appuyant sur deux enregistrements issus de l'étude de Frost et al. (2024). Ces enregistrements incluent un alignement entièrement vérifié, pouvant ainsi servir de référence. Comme expliqué dans la chapitre 5.1.2, deux métriques ont été calculées :

- **Le score de précision** (P), qui mesure la proportion de durée alignée automatiquement correspondant à l'alignement de référence ;
- **Le score de rappel** (R), qui mesure la proportion de l'alignement de référence correctement identifié par l'aligneur.

Nous souhaitons en particulier obtenir un score de précision élevé, car il est important que l'alignement corresponde précisément à ce qui est dit dans l'enregistrement, même si certaines portions ne sont pas alignées. Le score de rappel, quant à lui, renseigne sur la proportion de la référence qui a été correctement alignée.

La comparaison entre l'alignement de référence (REF) et l'alignement automatique (AUTO) s'effectue au niveau des mots (intervalles pleins), en excluant les intervalles vides. Pour chaque mot aligné automatiquement, nous avons mesuré la durée correspondant au même mot dans REF, représentée par un segment vert dans la figure 7.4.

Sur les 442 s d'enregistrement du corpus de référence, Wav2Vec a aligné 239 s, contre 315 s pour l'alignement de référence. Parmi ces 239 s alignées par Wav2Vec, 218 s correspondent effectivement à REF, ce qui donne un score de précision élevé ($P = 0{,}91$). Cependant, la figure 7.4 montre que les segments alignés par Wav2Vec sont souvent légèrement plus courts que ceux de REF, ce qui reflète une tendance de l'aligneur à réduire la durée des mots par rapport à un alignement manuel. Cette limitation se traduit par un score de rappel plus faible ($R = 0{,}69$).

[2] https://huggingface.co/openai/whisper-medium

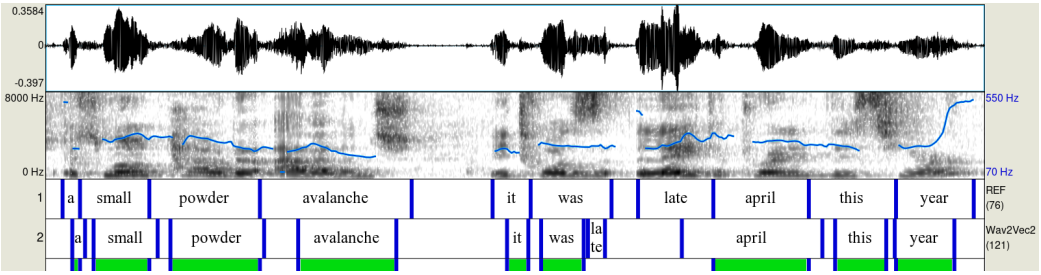


FIG. 7.4 : Visualisation de la correspondance entre l'alignement de référence (REF) et l'alignement automatique de Wav2Vec2.0. Les intervalles verts indiquent les portions d'alignement automatique correctes.

	Word Duration	Correct Duration	P	R
Reference	315	-	-	-
Wav2Vec 2.0	239	218	**0,91**	0,69
WebMaus 3.4	322	270	0,84	**0,86**
MFA 2.0	329	245	0,74	0,78

TAB. 7.4 : Scores de précision et de rappel obtenus par les différents systèmes d'alignement automatique évalués. Word Duration indique la durée totale d'alignement (en secondes).

Nous avons également évalué deux autres systèmes d'alignement sur le même corpus : WebMaus v3.4 et Montreal Forced Aligner v2.0. Ces systèmes présentent des scores de rappel plus élevés, respectivement $R = 0{,}86$ et $R = 0{,}78$, indiquant qu'une plus grande proportion de REF est alignée. Cependant, leurs scores de précision sont inférieurs, $P = 0{,}84$ pour WebMaus et $P = 0{,}74$ pour Montreal Forced Aligner, ce qui signifie qu'une part plus importante des alignements automatiques ne correspond pas à REF (voir tableau 7.4).

7.2 Annotation des pauses

L'évaluation de l'annotation des pauses a consisté en une comparaison entre les annotations automatiques produites par PLSPP et des annotations manuelles réalisées sur un corpus de parole présentant des caractéristiques similaires à celles des données analysées dans le cadre de notre étude sur la parole spontanée.

Ce corpus est composé de 72 dialogues spontanés enregistrés auprès de 24 binômes de locuteurs slovacophones (Mareková & Beňuš, 2024). Dans chaque dialogue, un des participants guide son interlocuteur en décrivant un itinéraire sur une carte, sans que les locuteurs puissent se voir. Les tours de parole sont transcrits et alignés au signal acoustique, et les pauses sont annotées selon plusieurs catégories : pauses

	N	d_{min}	d_{med}	d_{max}
btw	1804 (23,53%)	147	484	2470
mid	1937 (25,26%)	200	442	3548
fp	1416 (18,47%)	67	398	1492
p	2511 (32,75%)	11	319	2760
Total	**7668**			

Tab. 7.5 : *Nombre de pauses annotées manuellement présentes dans les 983 segments de parole extraits du corpus de Mareková et Beňuš (2024), et durée minimum, médiane et maximum (ms)*

inter-propositions (btw) et intra-propositions (mid), ainsi que pauses pleines (fp) et pauses inter-tours (p).

Nous avons analysé les 72 enregistrements avec l'ensemble des modules de PLSPP, incluant la segmentation en locuteurs, l'extraction des segments de parole, la transcription et l'alignement au signal, ainsi que les analyses syntaxiques et l'annotation des pauses. Ces processus ont permis d'extraire et d'analyser 983 segments. L'évaluation des annotations des pauses repose sur ces 983 segments.

Ces segments contiennent 7 668 pauses annotées manuellement[3]. Elles se répartissent comme suit : 24 % sont des pauses inter-propositions, 25 % intra-propositions, 18 % sont pleines, et 33 % inter-tours (*cf.* tableau 7.5). Il convient de noter que chaque pause est associée à une seule de ces catégories : les pauses pleines et inter-tours ne sont donc pas annotées en fonction de leur position syntaxique dans l'énoncé. Cependant, il est possible que deux pauses de types différents se succèdent, par exemple une pause pleine suivie d'une pause intra-propositionnelle.

L'annotation automatique réalisée par PLSPP a quant à elle identifié un total de 8 194 pauses d'une durée comprise entre 180 ms et 2 s[4]. La figure 7.5 illustre la comparaison des annotations. Les tiers 1 à 7 proviennent du corpus de Mareková et Beňuš (2024), tandis que la tier 8 représente la transcription alignée par PLSPP, où les pauses sont signalées en vert.

Pour chaque pause identifiée par PLSPP, nous avons calculé la proportion de chaque type de pause annotée manuellement sur la portion correspondante du signal. Par exemple, dans le cas illustré, la première pause détectée par PLSPP correspond à une pause inter-tour (p) pour 70 %, la deuxième ne correspond à aucune pause annotée manuellement, et la dernière correspond simultanément à deux pauses inter-

[3]Les pauses situées en début et en fin de segments ont été exclues de l'analyse, car elles peuvent avoir été tronquées lors de l'extraction.

[4]Les pauses situées en début et en fin de segments ont été ignorées ici également. Les seuils de durées minimum (180 ms) et maximum (2 s) ont été retenus car ils correspondent aux critères utilisés majoritairement dans les analyses présentées au chapitre suivant.

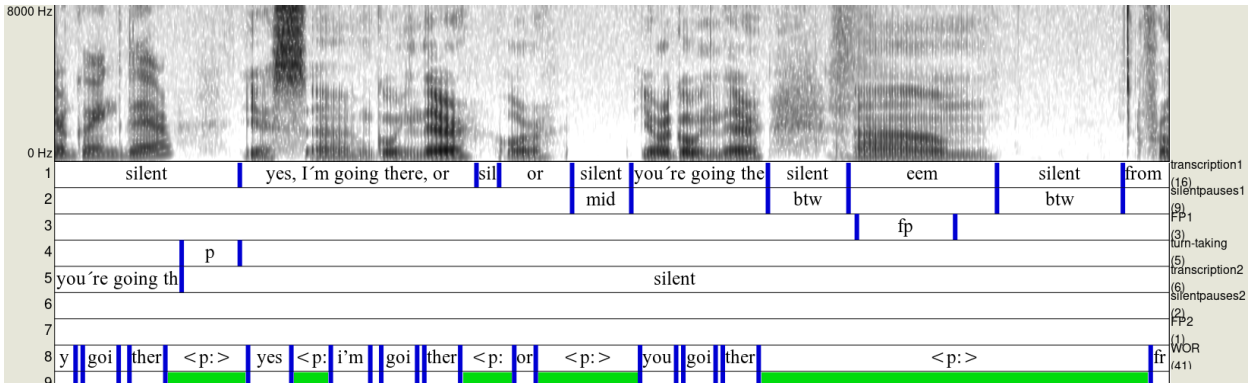


*FIG. 7.5 : Illustration de la comparaison de l'annotation des pauses entre PLSPP et l'annotation manuelle de Mareková et Beňuš (2024). La tier 1 correspond à la transcription manuelle du premier locuteur, la tier 2 aux pauses silencieuses (inter et intra-proposition), la tier 3 aux pauses pleines et la tier 4 aux pauses inter-tour. Les tiers 5, 6 et 7 correspondent aux 1, 2 et 3 pour le second locuteur. La tier 8 correspond à la transcription et l'alignement automatique de PLSPP. Les zones vertes indiquent les pauses identifiées par PLSPP (*180 ms-2 s*).*

	N	**btw**	**mid**	**fp**	**p**	**silent**	**Ø**
BC	2650	**662 (24,98%)**	207 (7,81%)	144 (5,43%)	**914 (34,49%)**	335 (12,64%)	388 (14,64%)
WC	5544	771 (13,91%)	**1149 (20,73%)**	424 (7,65%)	1001 (18,06%)	889 (16,04%)	**1310 (23,63%)**
Total	**8194**						

TAB. 7.6 : Type de pause annotée manuellement identifié pour chaque pause inter-proposition (BC) et intra-proposition (WC) annotée par PLSPP

propositions (btw) pour 53 % et à une pause pleine (fp) pour 25 %. À partir de ces proportions, nous avons déterminé le type de pause majoritaire pour chaque pause PLSPP, afin de faciliter la comparaison entre les annotations automatiques et manuelles.

La majorité des 2 650 pauses inter-propositions (BC) détectées par PLSPP correspondent à des pauses annotées comme inter-tours (p, 34 %), tandis que 25 % correspondent à des pauses inter-propositions (btw) et 8 % à des pauses intra-propositions (mid) (*cf.* tableau 7.6). Environ 15 % d'entre elles ne correspondent à aucune annotation manuelle (Ø), et 13 % correspondent à des intervalles notés "*silent*" sans annotation explicite de pause (comme illustré par la troisième pause PLSPP dans la figure 7.5).

Concernant les 5 544 pauses intra-propositions (WC) détectées par PLSPP, les correspondances avec les annotations manuelles sont les suivantes : 24 % ne correspondent à aucune annotation (Ø), 21 % à des pauses intra-propositions (mid), 18 %

à des pauses inter-tours (p), 16 % à des intervalles “silent”sans annotation explicite, 14 % à des pauses inter-propositions (btw), et 8 % à des pauses silencieuses (fp).

Ces résultats révèlent des performances mitigées. La faible précision des annotations automatiques pour les pauses inter- et intra-proposition (respectivement 25 % et 21 %) soulève plusieurs interrogations. Cette imprécision pourrait être liée à des erreurs dans l'étiquetage syntaxique, mais également au fait qu'un grand nombre de pauses sont annotées comme inter-tours dans le corpus de référence. Le choix du corpus de comparaison n'était peut-être pas optimal, dans la mesure où la présence de multiples catégories de pauses et d'un grand nombre de courtes réactions de l'interlocuteur, non éliminées malgré l'extraction des segments par PLSPP, pourrait compliquer la comparaison. Une évaluation spécifique de la précision de l'analyse grammaticale par constituant semble nécessaire pour mieux identifier les limites de l'annotation automatique des pauses.

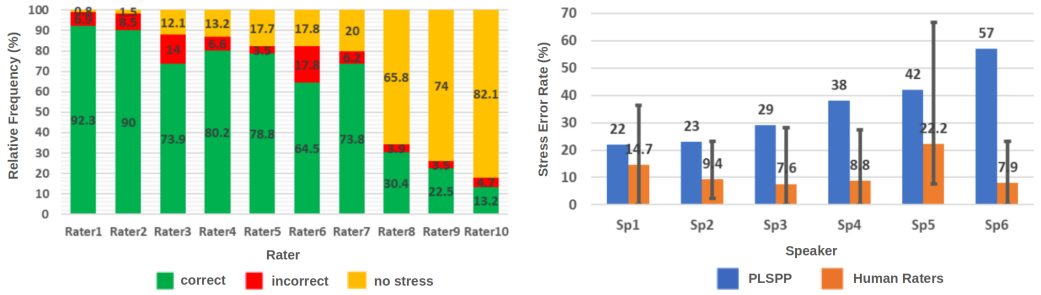
Enfin, le choix de n'inclure que les pauses PLSPP d'une durée comprise entre 180 ms et 2 s repose sur les critères retenus pour l'analyse des annotations des corpus CLES. Toutefois, un ajustement du seuil de durée minimum à 250 ms améliore légèrement les performances : la précision des pauses inter-proposition atteint alors 29 %, tandis que celle des pauses intra-proposition passe à 25 %. Par ailleurs, la proportion de pauses non annotées dans le corpus de référence diminue, atteignant 8 % et 13 % respectivement pour les pauses inter- et intra-proposition. On peut donc penser qu'un certain nombre de pauses courtes n'ont pas été annotées manuellement.

7.3 Annotation de l'accent lexical

3.1 Évaluation perceptive par des auditeurs natifs

Pour évaluer la qualité de l'annotation de l'accentuation lexicale par PLSPP, nous avons comparé les résultats obtenus avec ceux d'une annotation manuelle, réalisée par des locuteurs anglophones natifs. Nous présentons ici les résultats obtenus grâce au travail de recherche d'un étudiant du *Spoken Language Processing Laboratory* de l'université Dōshisha, détaillés plus longuement par Kimura et al. (2024).

Dix évaluateurs ont été recrutés pour annoter manuellement la syllabe qu'ils percevaient comme accentuée dans les mots polysyllabiques lexicaux d'un texte lu par six locuteurs anglophones de langue maternelle japonaise. Chaque mot cible a ensuite été classé comme « correct », « incorrect » ou « sans accent », selon les annotations des participants.



(a) Distribution des annotations de l'accent lexical par les 10 évaluateurs à travers les six enregistrements

(b) Taux d'erreur d'accentuation par locuteur selon PLSPP et les 10 évaluateurs humains. La barre verticale indique l'amplitude des scores humains

FIG. 7.6 : Figures issues de Kimura et al. (2024, p. 675)

La figure 7.6a présente la distribution des annotations par évaluateur. Une forte variabilité apparaît quant à la perception de la présence ou de l'absence d'un accent lexical. Les évaluateurs 8, 9 et 10 rapportent une absence d'accent pour la majorité des mots (respectivement 66, 74 et 82 %), tandis que cinq autres identifient un accent sur 80 à 88 % des mots, et les deux premiers annotateurs sur 98 et 99 %. Il semblerait que le seuil de sensibilité pour percevoir la position de l'accent est un paramètre fortement dépendant des évaluateurs. Par ailleurs, le taux d'erreur d'accentuation varie entre 3,5 et 17,8 % (moyenne : 7,56 %; écart-type : 4,8), révélant un certain désaccord entre les évaluateurs.

En parallèle, les mêmes enregistrements ont été annotés automatiquement avec PLSPP v2. Les modules de segmentation par locuteur et de reconnaissance de la parole ont été désactivés, et la transcription orthographique du texte a été directement fournie au module d'alignement. La figure 7.6b compare le taux d'erreur d'accentuation entre PLSPP et les annotations humaines. Seuls les mots pour lesquels la position de l'accent était précisée ont été inclus. Il apparaît que PLSPP rapporte systématiquement davantage d'erreurs que les annotateurs humains, particulièrement pour le locuteur 6. Deux facteurs peuvent expliquer cet écart : d'une part, des erreurs d'alignement, notamment dans l'enregistrement du locuteur 6, qui présente plus de disfluences que les autres ; d'autre part, le fait que PLSPP identifie systématiquement une syllabe proéminente, alors que les annotateurs humains peuvent indiquer l'absence d'accent (dans ce cas, le mot n'est pas comptabilisé comme incorrect). Cela conduit à un jugement plus strict du système automatique.

Pour réduire l'impact des erreurs d'alignement sur les résultats, un filtrage manuel a été effectué pour ne conserver que les mots correctement alignés. Le nombre de mots retenus par locuteur varie entre 33 et 51 (M = 43), pour un total de 258 mots.

	Sp1	Sp2	Sp3	Sp4	Sp5	Sp6
Mots annotés par PLSPP	55	48	49	34	45	46
Mots dont l'alignement est correct	51	45	48	33	44	37

TAB. 7.7 : Nombre de mots annotés par PLSPP et nombre de mots conservés pour les analyses pour chaque locuteur

Le tableau 7.7 présente le nombre de mots annotés par PLSPP et ceux retenus après filtrage manuel pour chaque locuteur.

La figure 7.7 est une projection de ces 258 mots en fonction du score de contraste prosodique (C') estimé par PLSPP, et la moyenne des annotations humaines. Le coefficient de corrélation entre les deux mesures est faible ($r = 0{,}29$). Si l'on divise la figure en quatre zones, on obtient la distribution suivante :

- (a) 52 mots (20%) sont évalués corrects par la majorité des évaluateurs mais obtiennent un contraste prosodique négatif d'après PLSPP ($C' \leq 0{,}5$) ;
- (b) 164 mots (64%) sont évalués corrects par la majorité des évaluateurs et obtiennent un contraste prosodique positif ($C' > 0{,}5$) ;
- (c) 25 mots (10%) sont évalués incorrects par la majorité des évaluateurs et obtiennent un contraste prosodique négatif ;
- (d) 17 mots (7%) sont évalués incorrects par la majorité des évaluateurs mais obtiennent un contraste prosodique positif.

Le taux d'accord entre la moyenne des évaluations humaines et les estimations automatiques est de 73,3 % (coefficient κ de Cohen : 0,27). Il apparaît que la majorité des mots (83,7 %) sont jugés corrects par la majorité des annotateurs, bien qu'un consensus total soit rare (seulement 8 mots, soit 3,1 %). De plus, de nombreux mots avec un contraste accentuel légèrement négatif (C' entre 0,4 et 0,5) sont perçus comme correctement accentués par les annotateurs, tandis que les mots avec des contrastes fortement négatifs, bien que rares, sont généralement jugés incorrects.

Ces observations mettent en lumière deux points importants : premièrement, l'accentuation d'un mot ne devrait pas être considérée comme un paramètre binaire (correct/incorrect), mais plutôt comme un continuum reposant sur le contraste accentuel entre syllabes. Un contraste plus élevé est en effet plus facilement perçu par les auditeurs. Deuxièmement, les auditeurs semblent souvent identifier un schéma accentuel correct même lorsque le contraste prosodique est faible ou légèrement négatif (mauvaise syllabe accentuée). Ces résultats corroborent ceux de van Leyden et van Heuven (1996) et Cooper et al. (2002), qui montrent que les locuteurs anglophones

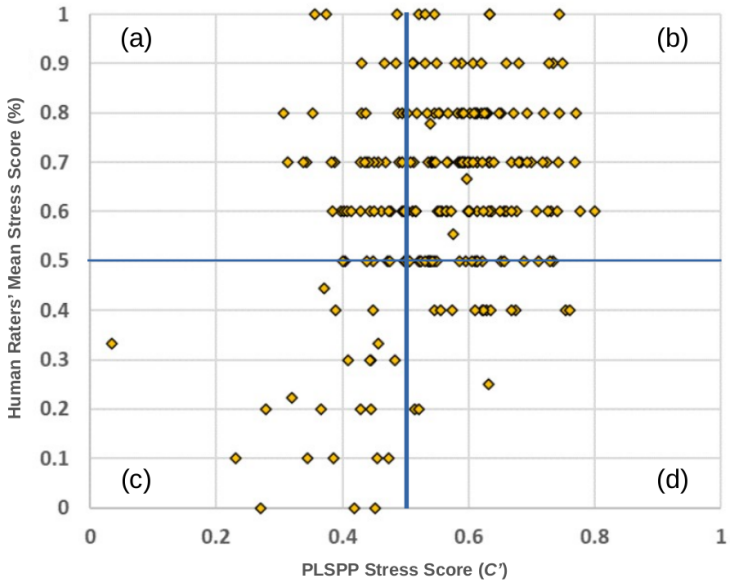


FIG. 7.7 : Score accentuel des 258 mots analysés : score estimé par PLSPP (C') en x et moyenne des évaluateurs humains en y (Kimura et al., 2024, p. 676). Un score de 0,5 en x indique un contraste prosodique nul entre les syllabes ; un score de 0,5 en y indique que 50 % des évaluateurs considèrent que l'accent est correctement positionné.

natifs ont tendance à percevoir un accent sur la syllabe initiale, même en l'absence d'indices lexicaux ou prosodiques. Cette tendance s'explique probablement par le fait que la majorité des mots usuels de l'anglais en parole spontanée sont accentués sur la syllabe initiale, et que ce pattern est donc plus « attendu » que les autres (Cutler & Carter, 1987).

Ainsi, cette étude montre que l'estimation automatique de l'accentuation lexicale par PLSPP est globalement en adéquation avec le jugement des auditeurs natifs, mais également que ce jugement humain varie de manière non négligeable selon les évaluateurs, qui semblent plus ou moins influencés par les tendances d'accentuation de leur langue maternelle. Cette variation inter-évaluateur souligne le fait que la perception de l'accent est un phénomène subjectif et contextuel, et que les paramètres prosodiques syllabiques ne font que participer à cette perception.

3.2 Annotation automatique et conscience phonologique

La deuxième investigation a consisté à confronter les schémas accentuels identifiés par PLSPP avec ceux dont les locuteurs ont conscience. Plus exactement, nous avons cherché à savoir si un locuteur produit effectivement une proéminence acous-

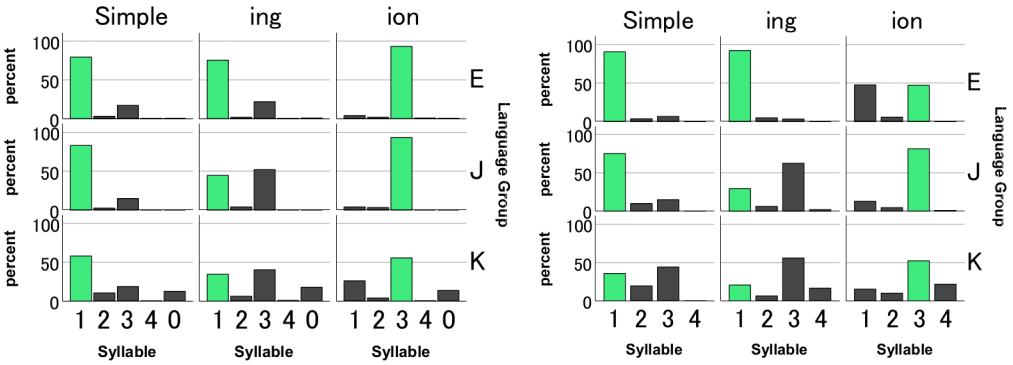
tique sur la syllabe qu'il pense devoir accentuer. Nous avons également investigué l'influence que peuvent avoir les tendances accentuelles de la langue maternelle (L1) des locuteur, en comparant trois groupes de locuteurs de L1 différentes. Cette section est un résumé d'une présentation donnée à la conférence LabPhon en 2024 (Sugahara et al., 2024).

Trois groupes de locuteurs ont donc participé à l'expérience : 12 locuteurs anglophones natifs (ENS), 14 locuteurs japonophones (JLE) et 11 locuteurs coréanophones (KLE) de niveaux CECRL B1 à B2. L'expérience a consisté en deux tâches : une annotation manuelle de la position de l'accent primaire sur une liste de mots cibles, via un questionnaire papier *(stress-assignment task)* – les participants devaient entourer la syllabe qui porte, selon eux, l'accent primaire –, et un enregistrement des mêmes mots cibles dans des phrases porteuses *(production task)*. Ces enregistrements ont ensuite été annotés automatiquement avec PLSPP v2 de la même manière que pour Kimura et al. (2024).

Les mots sélectionnés consistent en 19 triplets composés d'un verbe de 3 syllabes à l'infinitif portant l'accent sur l'initiale (ex. *dominate*), son participe présent (en *-ing*, accent primaire sur l'initiale, ex. *dominating*), et son dérivé substantif en *-ion* (accent primaire sur la 3ème syllabe, ex. *domination*). Pour plus de commodité, nous appellerons par la suite la première syllabe « $\sigma 1$ » et la troisième « $\sigma 3$ ».

À l'issue de la tâche d'annotation manuelle de l'accent, il est apparu que les ENS ont choisi majoritairement la position prescriptive de l'accent pour les trois formes (infinitif, participe présent et substantif). On note toutefois une certaine variabilité inter-annotateur pour l'infinitif et le participe (-ing), pour lesquels $\sigma 3$ a été identifiée comme portant l'accent primaire par un certain nombre de participants (*cf.* figure 7.8a, première ligne). De leur côté, les JLE ont choisi la position prescriptive pour l'infinitif et le substantif, mais se divisent en deux groupes pour le participe, $\sigma 3$ étant sélectionnée dans un peu plus de 50 % des cas (*cf.* figure 7.8a, deuxième ligne). Enfin, les KLE montrent des résultats plus variés, et ne choisissent parfois aucune syllabe (valeur 0, figure 7.8a, troisième ligne).

La figure 7.8b présente les résultats de l'estimation de la position de l'accent par PLSPP, et la figure 7.9b indique le taux d'accord par locuteur entre l'annotation et la production. On peut voir que $\sigma 1$ est clairement identifiée comme proéminente chez les ENS pour l'infinitif et le participe. Contrairement à la tâche d'annotation, il n'y a pas de variabilité inter-locuteur (proéminence exclusive sur $\sigma 1$). Du côté des substantifs, la proéminence est détectée tantôt sur $\sigma 1$ et $\sigma 3$, bien que $\sigma 3$ était exclusivement sélectionnée dans la tâche d'annotation. Bien que la proéminence soit détectée sur $\sigma 1$ pour une partie des substantifs, le poids de $\sigma 3$ (moyenne des trois dimensions) permet toutefois de distinguer les participes des substantifs (*cf.* figure 7.9a. Dans le cas du substantif, $\sigma 1$ porte l'accent secondaire, mais elle a peut-être tendance à être accentuée



(a) Résultats de l'annotation manuelle de la position de l'accent (stress-assignment task)

(b) Résultats de l'estimation de la position de l'accent par PLSPP (production task)

Fig. 7.8 : *Sur les deux figure, la barre verte représente la position prescriptive de l'accent primaire. Simple, ing et ion font référence à l'infinitif, au participe et au substantif; E, J et K aux locuteurs anglophones, japonophones et coréanophones. Le chiffre en x indique la position de la syllabe accentuée. Des figures plus détaillées sont présentées en annexe F.*

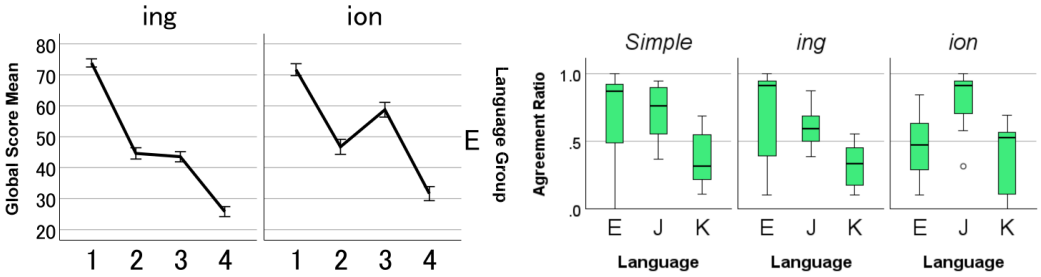
plus fortement par sous l'influence d'une tendance à accentuer la première syllabe en anglais. Il peut également s'agir d'un artefact provoqué par la phrase porteuse et le contexte très contraint de la production.

Les locuteurs JLE présentent de manière générale un haut degré de corrélation entre l'annotation de la position de l'accent et la détection automatique de la syllabe proéminente. Contrairement aux ENS, un contraste clair est observable sur $\sigma 3$ dans le cas du substantif. Cela pourrait peut-être s'expliquer par le fait que les japonais sont habitués accentuer en médiale dans leur langue maternelle, et à n'accentuer qu'une seule syllabe par mot.

Enfin, dans le cas des KLE, on observe un faible taux d'accord entre les annotations et la détection automatique de proéminence pour les trois formes, *cf.* figure 7.9b.

Le haut degré d'accord annotation-production chez les locuteurs ENS et JLE laisse penser que l'estimation de la position de l'accent par PLSPP est généralement correcte dans ce type de tâche contrôlée. Par ailleurs, l'estimation de position de l'accent basée sur la moyenne des trois dimensions prosodiques (f_0, intensité, durée) donne de meilleurs résultats que chaque dimension de manière isolée. Il semble donc important de considérer les trois dimensions pour estimer la position de l'accent. Les résultats de l'estimation automatique par dimension et le taux d'accord associé sont présentés en annexe F.

Nous avons constaté dans cette étude une influence de la L1 des locuteurs, tant au niveau de la tâche d'annotation que de la tâche de production. Si les ENS ont généralement choisi la position prescriptive de l'accent dans la tâche d'annotation, on observe



(a) Centile moyen par syllabe pour les participes (-ing) et les substantifs (-ion) dont la proéminence est détectée sur σ1 par PLSPP, chez les locuteurs ENS. Les barres d'erreur représentent l'intervalle de confiance à 95 %.

(b) Taux d'accord par locuteur entre l'annotation manuelle de la position de l'accent par le locuteur, et la détection de proéminence par PLSPP. E, J et K correspondent aux groupes de locuteurs anglophones, japonophones et coréanophones.

***Fig. 7.9 :** Pour plus de détails, se référer à l'annexe F*

une tendance à accentuer $\sigma 1$ même lorsque l'accent est conscientisé sur $\sigma 3$. Du côté des JLE, on constate une tendance à souvent accentuer $\sigma 3$ quel que soit le type de mot, cela pouvant être une influence de la plus fréquente accentuation en médiale en japonais. Enfin, les locuteurs KLE présentent effectivement plus de difficultés de manière générale pour placer l'accent, autant pour la tâche d'annotation que de production : le taux d'accord entre annotation et production est généralement en-dessous de 0,5. On constate que les locuteurs JLE obtiennent un taux d'accord annotation-production largement supérieurs au KLE, et qu'ils placent généralement l'accent sur la position prescrite. On peut y voir ici l'influence de la présence d'un accent lexical en japonais, contrairement au coréen, qui simplifie donc la conscientisation et la production de l'accent en anglais.

3.3 Annotation de parole produite par des locuteurs natifs

Dans cette troisième approche pour évaluer les performances de PLSPP en termes d'annotation de l'accentuation lexicale, nous avons analysé l'annotation automatique obtenue à partir de parole produite par des locuteur natifs. Ici, nous faisons le postulat que les locuteurs natifs ont généralement tendance à accentuer la syllabe prescrite, et qu'ils peuvent donc être considérés comme une référence. De cette manière, si la syllabe proéminente identifiée par PLSPP ne correspond pas à la syllabe censée porter l'accent primaire selon le dictionnaire phonologique intégré à l'outil, on peut considérer que l'annotation est incorrecte.

Deux corpus de parole contrôlée ont été analysés avec PLSPP v2 :

	Phrases porteuses		Textes lus	
	Filtré	*Non-filtré*	Filtré	*Non-filtré*
Nombre de mots polysyllabiques lexicaux annotés	541	*954*	4414	*7238*
Score de position de l'accent (S)	79 %	*76 %*	73 %	*71 %*
Contraste prosodique moyen ($\overline{C}$)	28	*27*	17	*16*
Contraste de f_0 ($\overline{C_{f_0}}$)	26	*25*	15	*15*
Contraste d'intensité ($\overline{C_{int}}$)	33	*33*	21	*21*
Contraste de durée ($\overline{C_{dur}}$)	25	*23*	13	*11*
Centile moyen de la syllabe accentuée ($\overline{P_s}$)	70	*70*	61	*60*

Tab. 7.8 : *Résultat des annotations automatiques obtenues sur de la parole native en production de parole contrôlée*

- L'enregistrement des 57 mots cibles dans des phrases porteuses lues par 17 locuteurs natifs (dont 12 sont issus de l'étude présentée dans la sous-section précédente) ;
- 92 textes lus en studio par 7 locuteurs natifs, issus de quatre manuels scolaires (Nakanishi et al., 2023a, 2023b, 2024a, 2024b). Les enregistrements vont de 55 s à 3 min 54 s (moyenne : 2 min 6 s), totalisant 3 h 13 min 18 s.

Un total de 954 mots polysyllabiques lexicaux ont été annotés pour le corpus de phrases porteuses, et 7 238 mots pour le corpus de textes lus. Pour réduire l'impacte de possibles erreurs d'alignement, nous avons filtré ces mots pour ne conserver que ceux présentant un nombre équivalent de syllabes et de pics d'intensité, ramenant le nombre de mots analysés à respectivement 541 et 4 414. Le tableau 7.8 présente les résultats obtenus avec et sans filtrage.

Dans le cas des phrases porteuses, 79 % des mots analysés sont accentués selon le dictionnaire de référence. Plus précisément, sur les trois types de mots produits, les infinitifs et les participes sont accentués en initiale pour 96 % chacun, tandis que les substantifs sont accentués en $\sigma 3$ seulement pour 46 % (contre 50 % en initiale). On retrouve le constat présenté dans la sous-section précédente, selon lequel les locuteurs natifs produisent une syllabe initiale marquée plus fortement pour la moitié des substantifs analysés.

La figure 7.10 montre le contraste moyen entre les syllabes des trois catégories de mots. Chaque syllabe est représentée par un cercle d'une taille proportionnelle à son poids moyen, c'est-à-dire à la moyenne des centiles de la syllabe en question sur l'ensemble des mots. Aussi, on peut voir que le contraste prosodique entre les syllabes est assez marqué, en particulier au niveau de la f_0 et de l'intensité. Dans le cas de la durée syllabique, $\sigma 3$ est relativement plus longue que les autres syllabes, probablement à cause de la présence d'une diphtongue dans la plupart des mots, dont la durée est

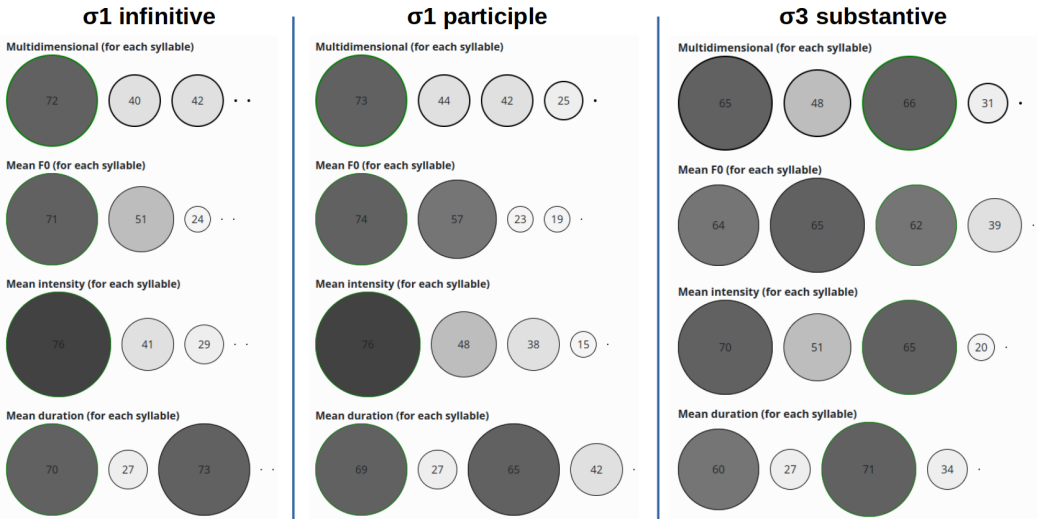


FIG. 7.10 : *Centile moyen par syllabe pour les trois types de mots produits dans les phrases porteuses : infinitifs (n=203), participes (n=155) et substantifs (n=183)*

intrinsèquement plus longue que les autres voyelles. Le cas du substantif est plus ambigu : $\sigma 1$ et $\sigma 3$ sont fortement marquées sur les trois dimensions prosodiques. Si on peut y voir là une proéminence due à l'accent secondaire, on ne voit pas le cas inverse pour le participe présent, ce qui laisse penser qu'il y a bien une préférence à accentuer la syllabe initiale, au moins dans le cadre de la production de ces phrases porteuses.

Dans le cas de la lecture de textes, le score de position de l'accent est légèrement moins élevé (73 %), de même que le contraste prosodique moyen entre les syllabes ($\overline{C} = 17$, contre 28 dans le cas des phrases porteuses). La figure 7.11 est une visualisation des différents patterns accentuels observés (syllabe proéminente identifiée par PLSPP) pour chaque gabarit accentuel du dictionnaire. Par exemple, les mots de gabarit Oo (mots de deux syllabes avec accent en initiale), comme "*student*" ou "*wonder*" , sont accentués selon PLSPP à 77 % en initiale et 23 % en finale. On peut voir que PLSPP détecte une proéminence sur la syllabe initiale pour une portion non négligeable de mots qui ne sont pas censés être accentués sur cette syllabe : 43 % de mots trisyllabiques censés porter l'accent en finale (ooO), 42 % des mots quadrisyllabiques censés le porter en $\sigma 3$. Dans le premier cas, une analyse approfondie du contraste syllabique par dimension des 87 mots de gabarit théorique ooO révèle que $\sigma 1$ a tendance à être allongée, bien que la proéminence soit détectée sur $\sigma 3$ (centile moyen de durée par syllabe : 56 22 77), et produite avec une intensité plus forte (64 37 52).

Ces résultats montrent que l'annotation de l'accentuation lexicale de PLSPP permet d'obtenir une représentation relativement fiable du contraste prosodique des syllabes, bien qu'elle reste influencée par la durée intrinsèque des voyelles. Il est diffi-

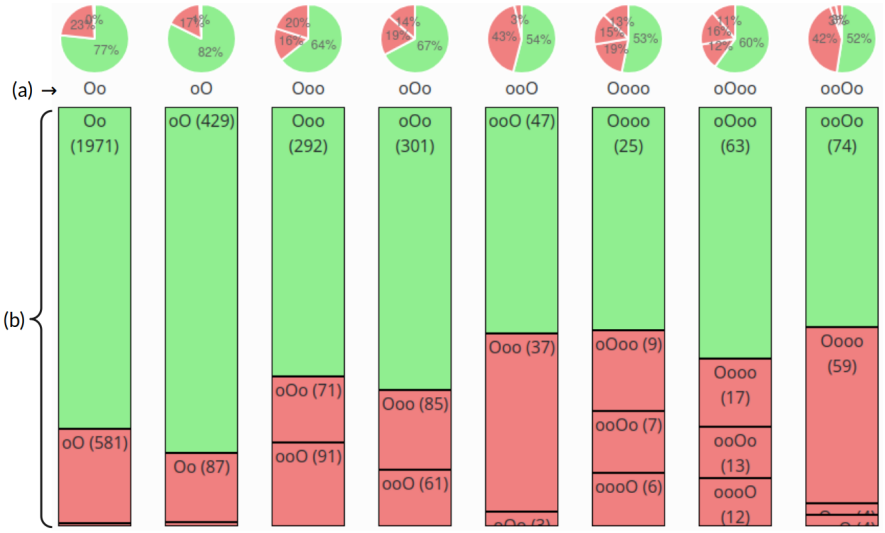


FIG. 7.11 : Position de la proéminence identifiée (b) pour chaque position théorique (a) parmi les 4 414 mots polysyllabiques lexicaux analysés dans les textes lus. "O" représente la syllabe accentuée d'après le dictionnaire (a) ou l'estimation de PLSPP (b). Le nombre entre parenthèses indique le nombre de mots.

cile d'identifier la cause précise des patterns accentuels identifiés comme incorrects par PLSPP : il peut s'agir d'erreurs de mesure, mais aussi d'une réalité acoustique dans la production des locuteurs natifs. Nous retiendrons notamment une tendance de PLSPP à mesurer une proéminence sur la syllabe initiale des mots.

Conclusion

Nous avons proposé une évaluation des différents modules de traitement, afin de vérifier la fiabilité des annotations produites et d'identifier les limites de l'outil. L'évaluation des modules de segmentation en locuteur, de reconnaissance de la parole et d'alignement ont permis de nous assurer que les étapes de prétraitements sont suffisamment performantes pour effectuer une annotation des pauses et de l'accentuation lexicale sur un corpus de conversations spontanées L2. On notera toutefois une précision limitée lorsque les chevauchements entre locuteurs sont nombreux, ainsi qu'une tendance générale de l'aligneur mot-signal à réduire légèrement la durée des mots. Aucune évaluation de l'analyse syntaxique n'a pu être effectuée faute de corpus de référence, mais elle semble nécessaire au vu des résultats mitigés obtenus lors de l'évaluation de l'annotation des pauses.

Enfin, l'évaluation de l'annotation de l'accentuation lexicale a mis en évidence le fait que l'accent est loin d'être un phénomène absolu et binaire (correct/incorrect), et

qu'il est préférable de considérer la mesure des proéminences acoustiques comme un degré de contraste continu entre les syllabes, participant à la perception de l'accent et plus généralement du rythme de la parole. Nous avons constaté que la perception de l'accent varie selon les auditeurs, et qu'elle subit probablement une influence des tendances accentuelles de leur langue maternelle. Cette influence semble également s'observer dans la production des locuteurs natifs, chez qui une importante proportion de mots accentués sur la syllabe initiale a été constatée.