Chapitre 9

Mesure de l'impact du rythme sur la compréhensibilité

Ce chapitre présente les résultats obtenus lors de l'expérience d'évaluation dynamique de la compréhensibilité que nous avons organisée en février 2024. Ces résultats ont d'abord été présentés lors du colloque English Pronunciation : Issues and Practice (EPIP8) en mai 2024 (Coulange et al., 2024b), puis ont fait l'objet d'une publication présentée à InterSpeech 2024 (Coulange et al., 2024c). Une analyse approfondie des 800 commentaires libres recueillis durant l'expérience est en cours et sera présentée prochainement, à la conférence New Sounds 2025 (Nakanishi & Coulange, 2025). Dans ce chapitre, nous présentons d'abord le comportement des évaluateurs qui ont participé à l'expérience, avant d'exposer les résultats des évaluations globales des enregistrements, puis l'analyse de l'évaluation dynamique de la compréhensibilité.

9.1 Comportements des évaluateurs

L'évaluation des 16 stimuli audio a duré en moyenne 26 min 59 s pour les 60 participants, allant de 12 min 42 s à 1 h 3 min 2 s, dont 4 évaluateurs excédant 45 min. Comme prévu, une grande variabilité de comportement a été observée parmi les évaluateurs, avec un nombre total de clics par évaluateur allant de 12 à 272 sur les 16 enregistrements (moyenne de 76,7, écart type de 48,65). Cinq évaluateurs ont présenté une fréquence de clics particulièrement élevée, totalisant plus de 120 clics chacun.

Si la fréquence de clic varie d'un participant à l'autre, une tendance claire à cliquer dans les mêmes zones est toutefois observable, se traduisant par des pics de clics relativement bien contrastés comme l'illustrent les figures 9.1 et 9.2.

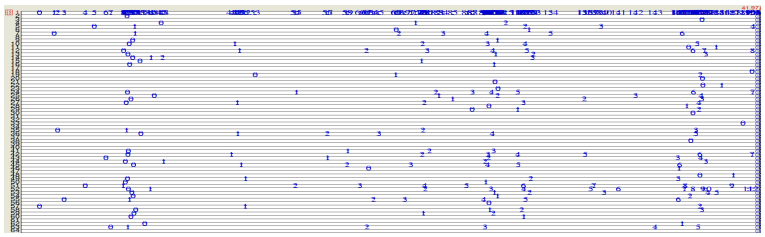

FIG. 9.1 : Aperçu des clics enregistrés par les 60 participants sur l'un des stimuli (format TextGrid, un point représente un clic, un évaluateur par tier, la première tier est la somme de l'ensemble des clics)

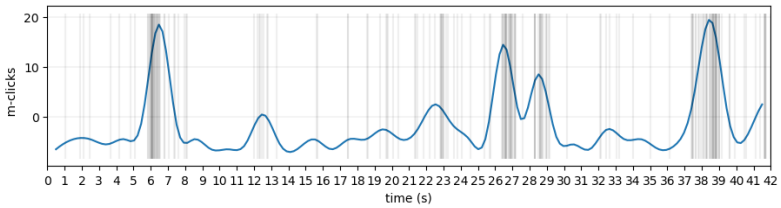

FIG. 9.2 : Somme des m-clics sur une fenêtre glissante d'une seconde, pour le même enregistrement (les clics bruts sont représentés par une barre verticale grise)

Le coefficient de corrélation intra-classe pour l'accord absolu entre les 60 évaluateurs (ICC(2, 60)) est de 0,97, avec un intervalle de confiance à 95 % de $[0{,}95; 0{,}99]$. Cet ICC élevé indique une forte cohérence entre les évaluateurs ($F(55, 15) = 55$, $p < 0{,}001$), confirmant la fiabilité des évaluations sur les dimensions évaluées. L'alpha de Cronbach calculé pour évaluer la cohérence interne des 60 évaluations sur les trois dimensions est de 0,93 ($IC = [0{,}93; 0{,}94]$) et indique une forte fiabilité des évaluations. Si l'une des dimensions était retirée, l'alpha de Cronbach resterait au-dessus de 0,89 pour chacune, ce qui montre que chaque dimension contribue de manière significative à la cohérence interne globale des évaluations. Les coefficients de corrélation item-total pour chaque dimension sont également élevés (prononciation : 0,86 ; fluidité : 0,85 ; compréhensibilité : 0,88), ce qui montre que chaque dimension est bien corrélée avec l'ensemble des évaluations. Ces valeurs sont proches de celles obtenues dans des études antérieures, comme celle de Kahng (2018) qui avait un accord absolu de 0,93 et un alpha de Cronbach de 0,98 dans une expérience mobilisant 46 évaluateurs et 80 extraits de parole. Ces mesures renforcent la validité des scores obtenus, indiquant que les différences observées entre les enregistrements reflètent bien des différences réelles, et non des variations inter- ou intra-individuelles dans le jugement des évaluateurs.

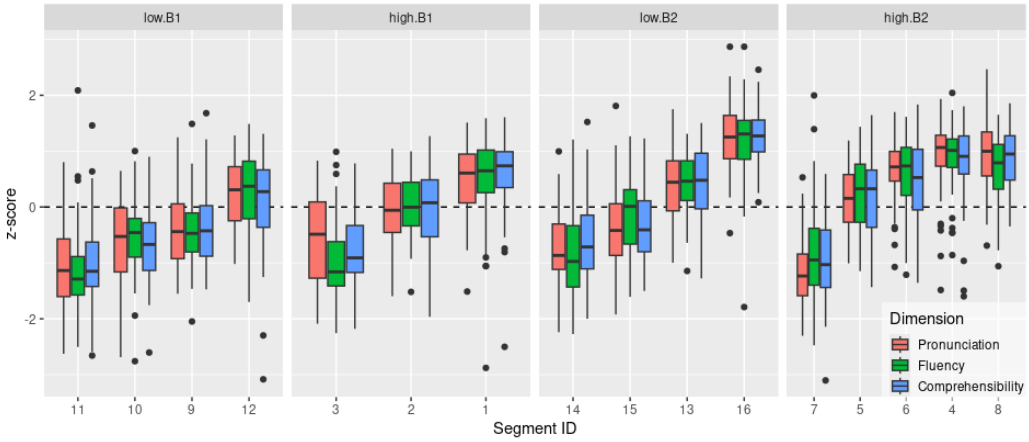

FIG. 9.3 : Évaluation globale normalisée des 16 segments en termes de qualité de prononciation (rouge) de fluidité (vert) et de compréhensibilité (bleu), en fonction du niveau du locuteur et de la catégorie du segment (low : haute proportion de pauses WP et bas score accentuel moyen ; high : basse proportion de pauses WP et haut score accentuel moyen ; un point de donnée correspond à l'évaluation d'un segment par un évaluateur sur une dimension)

9.2 Évaluations globales

Commençons par analyser les évaluations globales des enregistrements. Les scores de qualité de prononciation, de fluidité et de compréhensibilité apparaissent de manière générale très corrélés entre eux, comme le montre la figure 9.3. On peut voir que les locuteurs B2 ont tendance à obtenir un score global plus élevé que les B1 ($p < 0{,}001$, médianes à -0,31 pour B1 et +0,38 pour B2, $\Delta = -0{,}328$ (faible), $IC = [-0{,}367; -0{,}288]$, figure 9.4 gauche), bien que tous ne reçoivent pas un score positif. De la même façon, les segments catégorisés "high" par PLSPP (c'est à dire avec proportionnellement peu de pauses de type intra-syntagme et un score accentuel élevé) reçoivent un score généralement plus élevé que les segments "low", bien que le contraste soit moins important que pour le niveau du locuteur ($p < 0{,}001$, médianes à -0,22 pour low et +0,35 pour high, $\Delta = -0{,}202$ (faible), $IC = [-0.243; -0.16]$, figure 9.4 milieu). Notons par ailleurs que les locutrices ont tendance à obtenir un meilleur score ($p < 0{,}001$, médianes à -0,28 pour les hommes et +0,28 pour les femmes, $\Delta = -0{,}245$ (faible), $IC = [-0{,}286; -0{,}204]$, figure 9.4 droite).

Nous avons ensuite regroupé les segments en fonction de leurs tendances pour chaque dimension : ceux qui ont plus de pauses intra-syntagme, ceux qui ont les meilleurs scores accentuels etc.

Commençons par les pauses. On constate tout d'abord que les segments qui contiennent globalement moins de pauses de manière générale (les 8 segments dont

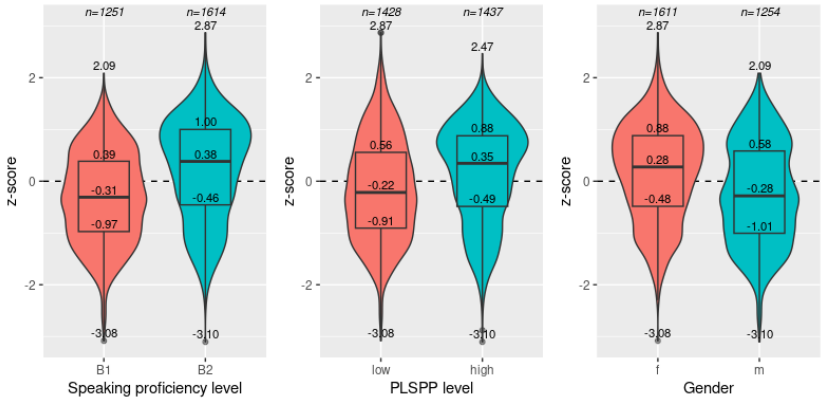

FIG. 9.4 : Évaluation globale normalisée en fonction du niveau, de la catégorie et du genre du locuteur (3 dimensions confondues, un point de donnée correspond à l'évaluation d'un segment par un évaluateur sur une dimension)

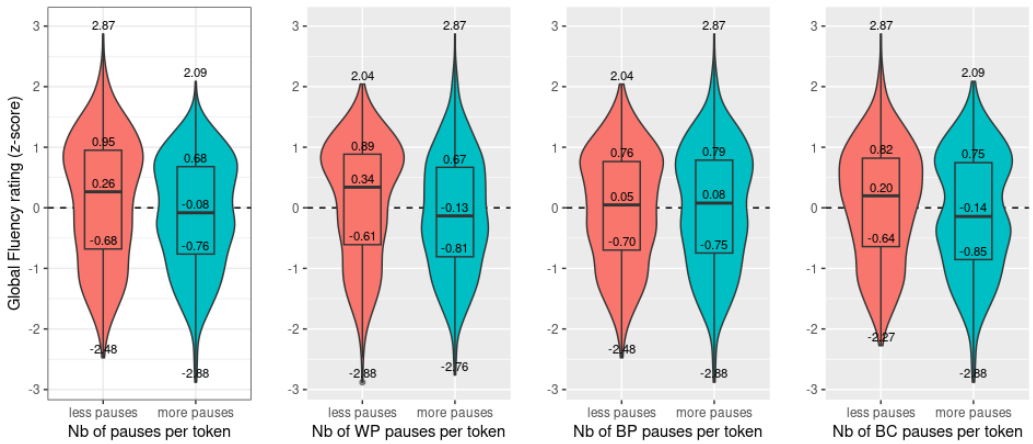

FIG. 9.5 : Évaluation globale de la fluidité selon la fréquence des différents types de pauses

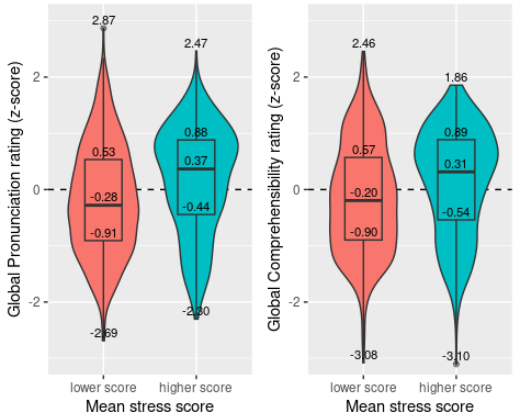

FIG. 9.6 : Évaluation globale de la prononciation et de la compréhensibilité en fonction du score accentuel moyen

le nombre de pauses par token est inférieur à la fréquence médiane) obtiennent un meilleur score de fluidité que les segments qui contiennent plus de pauses ($p < 0{,}001$, médianes à +0,26 contre -0,08, $\Delta = -0{,}138$ (faible), $IC = [-0{,}063; -0.212]$, *cf.* figure 9.5). Ce n'est pas une surprise : les pauses sont souvent perçues comme une disfluence de la parole, il n'est donc pas étonnant que les enregistrements qui en présentent plus soient jugés moins fluides. Mais regardons ce qu'il en est si l'on considère les pauses en fonction de leur position syntaxique : seules les pauses intra-syntagme (WP) présentent une différence significative sur la perception de fluidité des segments ($p < 0{,}001$, médianes à +0,34 contre -0,13, $\Delta = 0{,}156$ (faible), $IC = [0{,}082; 0{,}227]$) ; tandis que la fréquence des pauses inter-syntagme (BP) ne distinguent pas les segments (*ns.*, médianes à 0,05 et 0,08, $\Delta = -0{,}003$ (négligible), $IC = [-0{,}076; 0.07]$), et que celle des pauses inter-proposition (BC) reste assez floue : les segments qui en ont moins ont tendance à être légèrement mieux notés, mais la taille de l'effet reste négligeable ($p < 0{,}05$, médianes à +0,2 contre -0,14, $\Delta = 0{,}088$ (négligible), $IC = [0{,}014; 0{,}16]$). On observe les mêmes tendances avec le jugement de compréhensibilité.

Du côté de l'accentuation lexicale, on constate que les segments dont le score accentuel est bas ont tendance à être moins bien notés en termes de prononciation ($p < 0{,}001$, médianes à -0,28 contre +0,37, $\Delta = -0{,}225$ (faible), $IC = [-0{,}296; -0{,}153]$) comme de compréhensibilité ($p < 0{,}001$, médianes à -0,20 contre +0,31, $\Delta = -0{,}189$ (faible), $IC = [-0{,}26; -0{,}115]$, *cf.* figures 9.6).

Enfin, nous nous sommes intéressés au nombre total de clics par enregistrement. La figure 9.7 présente le nombre de clics normalisés reçus pour chaque segment audio en fonction du niveau du locuteur et de la catégorie du segment. La différence entre les locuteurs B1 et B2 apparaît très clairement ($p < 0{,}05$, médianes à 1,9 pour B1 et -2,1

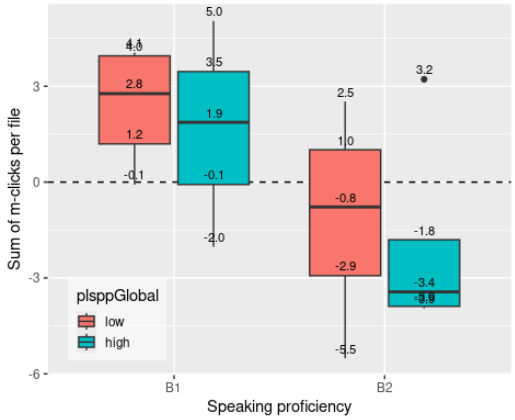

FIG. 9.7 : Somme des clics normalisés par enregistrement, groupés par niveau du locuteur et catégorie de segment

pour B2, $\Delta = 0{,}651$ (large) $IC = [0{,}088; 0{,}899]$). Les enregistrements catégorisés *low* obtiennent également plus de clics en moyenne que ceux catégorisés *high*, mais la différence n'est pas significative (médianes à 1,1 pour low et -1,9 pour high, $\Delta = -0{,}188$ (faible), $IC = [-0{,}679; 0{,}42]$).

À ce stade, nous avons vu que les enregistrements qui contiennent plus de pauses en général obtiennent de moins bons scores de fluidité et de compréhensibilité, mais que cette différence n'est significative que lorsqu'on regroupe les segments par fréquence de pauses intra-syntagme ; un plus grand nombre de pauses inter-syntagme ou inter-proposition n'affecte pas autant le jugement des évaluateurs. Nous avons vu également que les segments au score accentuel élevé obtiennent de meilleurs scores de prononciation et de compréhensibilité. Toutefois, s'il y a corrélation, il n'y a pas nécessairement causalité : on ne peut pas affirmer que les pauses ou les mots mal accentués ont un impact direct sur la compréhensibilité. Une évaluation dynamique de la compréhensibilité nous permettra d'observer les tendances de fluctuation du jugement à la suite de ces phénomènes, et ainsi mieux comprendre leur impact sur la perception de la parole.

9.3 Analyse des patterns de clics

Nous présentons dans cette section les résultats de l'analyse des variations du nombre de clics normalisé (m-clics) à la suite des pauses et des mots polysyllabiques. Nous utilisons une fenêtre glissante d'une seconde sur les 5 secondes suivant l'onset de l'événement qui nous intéresse. La moyenne des m-clics est calculée dans chaque fenêtre, et comparée selon le type de pause ou de pattern accentuel.

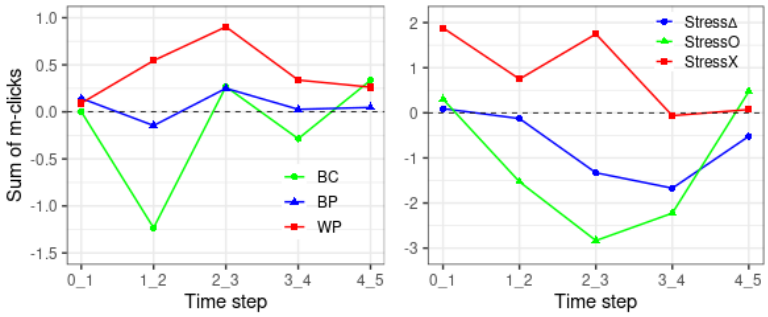

Fig. 9.8 : *Nombre de m-clics moyen sur les 5 secondes suivant l'onset d'une pause (gauche) ou d'un mot pattern accentuel (droite) ; les valeurs positives indiquent une activité de clics supérieure à la moyenne*

	Rank tests		*Pearson correlations*	
	BC vs. WP	**StressO vs. StressX**	**Stress score**	
window	p-value	p-value	R	p-value
0-1s	—	—	-0.13	—
1-2s	*	*	-0.1	—
2-3s	—	**	-0.25	**
3-4s	—	*	-0.062	—
4-5s	—	—	-0.027	—

Tab. 9.1 : *Tests de rangs comparant le nombre moyen de m-clics après les pauses inter-proposition (BC) et intra-syntagme (WP), et après les patterns accentuels corrects (StressO, $C'' > 0{,}2$) et incorrects (StressX, $C'' < -0{,}2$), et coefficient de corrélation entre le nombre de m-clics et la valeur de C'' (— : non significatif, * : $p < .05$, ** : $p < .01$)*

La figure 9.8 (gauche) montre le nombre moyen de m-clics sur les 5 secondes suivant l'onset d'une pause. Les valeurs positives indiquent une activité de clic supérieure à la moyenne. Entre 0 et 1 seconde après l'onset de la pause, le nombre de m-clics est proche de 0 pour tous les types de pauses, indiquant que l'activité de clics est normale. À partir d'une seconde après l'onset, on constate que le nombre de clics tend à augmenter lorsqu'il s'agit d'une pause intra-syntagme (WP), atteignant son maximum (+0,8) entre 2 et 3 secondes après l'onset de la pause. Parallèlement, le nombre de m-clics à la suite d'une pause inter-proposition (BC) décroît nettement (-1,23) entre 1 et 2 secondes, puis revient rapidement vers 0 dès la troisième fenêtre. Dans le cas des pauses inter-syntagme (BP), enfin, le nombre de m-clics semble stagner autour de 0, n'indiquant aucune variation observable de l'activité. Le test de rangs montre une différence significative entre le nombre moyen de m-clics après les pauses WP et BC seulement sur la deuxième fenêtre, entre 1 et 2 secondes, *cf.* tableau 9.1.

En ce qui concerne l'évolution du nombre de clics à la suite des 139 mots polysyllabiques cibles, on constate une tendance assez similaire. Bien que la moyenne de

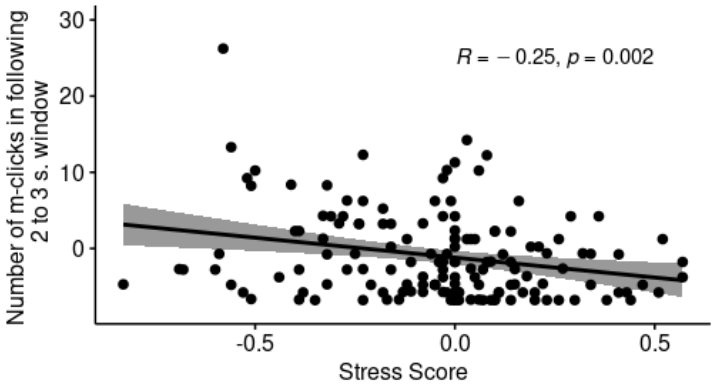

FIG. 9.9 : Projection des 139 mots cibles en fonction de leur score accentuel C'' et du nombre de m-clics enregistrés dans la fenêtre de 2 à 3 secondes après l'onset du mot

m-clics après les mots dont le pattern accentuel est jugé incorrect par PLSPP (StressX, $C'' < -0{,}2$) reste globalement supérieure à celle des mots jugés corrects (StressO, $C'' > 0{,}2$) ou ambigus (StressΔ), on observe une augmentation locale entre 2 et 3 secondes après l'onset du mot, mais une diminution évidente des clics après les mots StressO jusqu'à la troisième seconde (atteignant -2,83). La différence de nombre de m-clics après StressX et StressO est significative entre 1 et 4 secondes après l'onset du mot, *cf.* tableau 9.1.

Comme le score accentuel C'' est une valeur continue, nous avons également mesuré la corrélation entre celui-ci et le nombre de m-clics observés dans chaque fenêtre. La corrélation est négative de 0 à 5 secondes après l'onset, indiquant que plus le score est élevé, moins on observe de clics. La corrélation la plus forte, et la seule qui est significative, est observée entre 2 et 3 secondes : elle reste toutefois relativement faible (-0,25, $p < 0{,}01$, *cf.* tableau 9.1 et figure 9.9).

Conclusion

Le protocole expérimental mis en place dans cette étude a montré que les auditeurs natifs ont tendance à signaler des difficultés de compréhension dans les mêmes zones et que l'évaluation dynamique permet d'identifier ces zones et ainsi quantifier l'impact potentiel de certains phénomènes sur la compréhensibilité du locuteur.

Nous nous sommes concentrés sur l'impact des pauses et des patterns accentuels. Les analyses montrent une augmentation moyenne du nombre de clics sur les trois secondes suivant les pauses intra-syntagme, mais une diminution à la suite des pauses inter-proposition. Les enregistrements présentant plus de pauses en général ont été

moins bien notés en termes de fluidité et de compréhensibilité globale ; mais seule la proportion de pauses intra-syntagme semble faire la différence. Nous avons également mesuré une augmentation significative du nombre de clics entre deux et trois secondes après un mot au pattern accentuel incorrect, et à l'inverse une diminution sur trois secondes après un mot accentué sur la bonne syllabe. Ces résultats étaient attendus, puisque plusieurs études ont déjà démontré la corrélation entre la distribution des pauses et des accents lexicaux, mais notre approche propose un moyen de mesurer leur impact en temps réel sur la perception des auditeurs.