Chapitre 1

Évaluation de la prononciation

Dans une étude réalisée auprès de 459 enseignants d'anglais dans sept pays européens, Henderson et al. (2012) constatent que l'enseignement de la prononciation est souvent négligé en classe comme en formation de formateurs, amenant de grosses disparités dans la façon dont celle-ci est évaluée. Les enseignants disposent rarement des outils et de la formation nécessaires pour évaluer la prononciation de manière précise et systématique. Il en résulte souvent des « grilles d'évaluation maison » créées par les enseignants, et une importante variabilité dans les notes obtenues par les étudiants en fonction des évaluateurs (Frost & O'Donnell, 2018, p. 9). Gilquin et al. (2022) observent par exemple que les évaluateurs ne se basent pas sur les mêmes critères lorsqu'ils sont anglophones natifs ou non-natifs, ces derniers ayant tendance à juger de manière plus sévère, et à être plus souvent influencés par d'autres composantes de la production orale, comme la précision lexicale ou grammaticale. De leur côté, les évaluateurs natifs semblent donner une plus grande importance à l'intelligibilité globale du locuteur. Le degré de familiarité entre l'évaluateur et la prononciation du locuteur peut également avoir une influence sur l'évaluation : un évaluateur habitué à entendre un anglais produit par des locuteurs japonais aura souvent moins de mal à les comprendre qu'un évaluateur qui n'y est pas accoutumé (Didelot et al., 2019 ; Kim & di Gennaro, 2012 ; Minematsu et al., 2004). Ce manque de formation et d'outillage des enseignants pour évaluer la prononciation est relevé depuis de nombreuses années en Europe comme en Amérique du Nord (Baker, 2011 ; Burgess & Spencer, 2000 ; Derwing & Munro, 2015 ; Piccardo, 2016).

Gilquin et al. (2022) et Henderson et al. (2012) montrent un décalage important entre l'enseignement délivré en classe et les exigences des principaux tests certificatifs, pour lesquels la prononciation est un critère d'évaluation à part entière. Mais comment la prononciation est-elle évaluée dans ces tests certificatifs ?

Nous proposons d'examiner les descripteurs de compétences et les grilles d'évaluation de la prononciation que proposent les principaux tests certificatifs de l'anglais. Nous devrions ainsi mieux comprendre ce qu'il est attendu des apprenants à différents niveaux de compétence en langue, et identifier les bases sur lesquelles reposent (ou sont censées reposer) les jugements des évaluateurs. Nous nous intéresserons ensuite aux techniques d'évaluation automatique de la prononciation, en examinant les critères sur lesquels se basent les systèmes d'aujourd'hui, et en quoi ils diffèrent des évaluations humaines.

1.1 Évaluation humaine

1.1.1 Descripteurs du CECRL

Le Cadre Européen Commun de Référence pour les Langues (CECRL) est une initiative du Conseil de l'Europe pour définir des descripteurs de compétences détaillés afin de faciliter l'enseignement et l'évaluation des langues étrangères. La première édition de l'ouvrage (Conseil de l'Europe, 2001) propose une échelle dédiée à la « Maîtrise du système phonologique ». Cette échelle a toutefois été largement critiquée pour son manque de précision et ses descripteurs vagues basés sur l'intuition de l'évaluateur (ex. « net accent étranger » en A2, « prononciation clairement intelligible » en B1) et prenait pour modèle la prononciation d'un locuteur natif sans pour autant la définir (« prononciation et intonation claires et naturelles » au niveau B2). L'échelle complète est donnée en Annexe A.1.

Lors de la mise au point de la nouvelle édition des descripteurs en 2018, ces limitations ont été reconnues par les concepteurs des nouveaux descripteurs (Piccardo, 2016) qui qualifient ces descripteurs phonologiques d'« échelle la moins réussie » du CECRL et de seule échelle avec une norme native (p. 133). Les nouveaux descripteurs abandonnent la comparaison au modèle natif et se focalisent sur l'intelligibilité comme base théorique principale du contrôle phonologique. Les auteurs définissent l'intelligibilité comme « l'accessibilité du sens pour les auditeurs, incluant également la difficulté de compréhension perçue par les auditeurs (habituellement désignée comme compréhensibilité) » (Conseil de l'Europe, 2018, p. 140). On accepte maintenant un accent qui n'affecte pas la compréhension au niveau C2, ainsi que l'influence d'autre langues connues par l'apprenant. Il y a maintenant trois échelles : « Maîtrise générale du système phonologique », « Articulation des sons » et « Traits prosodiques » (Conseil de l'Europe, 2018, p. 142). Le tableau complet est donné en Annexe A.2.

Des termes relatifs à la prosodie, comme l'accent ou le rythme, sont maintenant mentionnés dès le niveau A1 : « très forte influence de l'accent, du rythme, et/ou de

l'intonation de l'une ou l'autre des langues qu'il parle » ; au niveau B1 « l'intonation et l'accentuation des énoncés et des mots sont presque corrects » ; au niveau B2 « peut en général [...] placer correctement l'accent », « l'accent a tendance à subir l'influence de l'une ou l'autre des langues qu'il/elle parle, mais l'impact sur la compréhension est négligeable ou nul » ; au niveau C1 « peut prononcer un discours fluide et intelligible en ne faisant que de rares erreurs d'accent, de rythme et/ou d'intonation qui n'affectent ni la compréhension ni l'efficacité ». Du côté de la prononciation des phonèmes, il est fait mention de « produire correctement des sons dans la langue cible » (A1), de prononciation « en général intelligible » (A2), mais aussi de « mauvaise prononciation systématique des phonèmes » (A2) ou des « erreurs de prononciation de sons et de mots » (B1, B2) sans plus de détails, laissant une part importante à l'interprétation de l'évaluateur. Notons qu'à partir du niveau B2, l'influence des caractéristiques phonologiques sur la compréhension devient « négligeable ou nul », et que le locuteur devient capable de « prédire avec une certaine précision les traits phonologiques de la plupart des mots non familiers (par ex. l'accent tonique en lisant) ».

Si la première édition du CECRL restait limitée au niveau de la prononciation, la nouvelle édition présente quant à elle des descripteurs plus détaillés, séparant la réalisation des phonèmes et les aspects prosodiques. Bien qu'ils ne soient pas exempts de critiques[1], ces descripteurs apportent déjà une base commune et solide pour évaluer la prononciation des apprenants.

1.1.2 Descripteurs du CLES

Le Certificat de Compétences en Langues de l'Enseignement Supérieur (CLES) est un test certificatif universitaire français établi par le Ministère de l'Enseignement Supérieur et de la Recherche. Le CLES est déployé aujourd'hui en 10 langues et proposé par une trentaine de centres CLES accrédités en France (rapport d'activité 2023[2]). Chaque niveau du CECRL est évalué indépendamment : le candidat doit choisir un niveau cible à valider lors de la passation de l'examen. Il existe des sessions CLES pour les niveaux B1, B2 et C1. Le CLES évalue quatre habiletés : la compréhension de l'écrit, la compréhension de l'oral, l'expression écrite ainsi que l'expression orale en monologue pour le niveau B1, et en interaction pour les niveaux B2 et C1.

Au niveau B2, l'examen consiste en une interaction orale sous la forme d'un jeu de rôle d'une dizaine de minutes à deux ou trois participants. Chaque participant se

[1] Didelot et al. (2019) critiquent notamment l'absence de considération des représentations sociales de l'auditeur sur la perception de l'intelligibilité, et le fait que le point de vue de l'auditeur de manière générale est peu pris en compte.

[2] Disponible en ligne à l'adresse suivante : https://www.certification-cles.fr/medias/fichier/rapport-d-activite-2023-certification-cles_1705953556233-pdf

voit attribuer un rôle en faveur ou contre un sujet polémique, comme l'usage de la cigarette électronique ou des tests cliniques sur les animaux par exemple. Les candidats disposent de deux minutes de préparation avant la discussion, puis doivent échanger leurs points de vue et argumenter pour arriver à un compromis dans un temps imparti de dix minutes. Ils sont évalués en direct par un ou deux évaluateurs accrédités présents dans la salle. L'évaluation est faite sur huit critères : la capacité à prendre position et négocier, la pertinence et la variété des arguments, la capacité à interagir, l'aisance, la phonologie, la cohérence du discours, la précision grammaticale et enfin la pertinence et la variété lexicale (*cf.* grille d'évaluation en Annexe A.7). Pour chacun des critères, l'évaluateur peut attribuer le niveau B2, ou à défaut B1 ou « non validé ». Le niveau B2 en interaction orale n'est validé que si l'ensemble des huit critères sont validés au niveau B2.

Concentrons-nous sur les deux critères qui relèvent de l'évaluation de la prononciation : l'aisance et la phonologie. Le premier fait référence à la capacité de l'étudiant à « exprimer ses idées avec fluidité sans faire de longues pauses (hésitations tolérées) », et « exprimer ses idées malgré des pauses pour chercher ses mots ». Le critère phonologie est décrit par une « prononciation et intonation suffisamment claire pour être aisément compris(e), même si un accent subsiste » et « globalement compréhensible malgré l'accent étranger et/ou des erreurs de prononciation ».

Sur le site du CLES, on peut lire qu'il est attendu du candidat de niveau B2 qu'il soit « significativement plus fluide et fasse moins d'erreurs » qu'au niveau B1, et soit « aisément compréhensible »[3]. Le niveau B2 semble donc caractérisé par une certaine fluidité de parole et d'aisance de compréhension côté auditeur.

Le CECRL et le CLES proposent des descripteurs communs à toutes les langues, mais qu'en est-il pour les descripteurs spécifiquement rédigés pour l'anglais L2 ?

1.1.3 Descripteurs du TOEFL

Le *Test of English as a Foreign Language* (TOEFL) est un test certificatif pour évaluer l'anglais langue seconde et développé par l'organisme privé *Educational Testing Service*. Il se décline en plusieurs versions adaptées à des publics allant du primaire à l'université. Nous nous intéresserons ici à deux de ces versions : le *TOEFL iBT*, qui évalue les compétences de l'apprenant en situation académique et qui est le test le plus répandu, et le *TOEFL ITP Assessment Series*, présenté comme un test à visée formative utilisé par certaines universités pour mieux adapter les enseignements aux besoins des apprenants. Les deux versions se passent sur ordinateur.

[3] https://www.certification-cles.fr/se-preparer/grilles-d-evaluation/grilles-d-evaluation-1196363.kjsp, consulté le 24/11/2024)

TOEFL iBT

Le TOEFL iBT met en avant l'évaluation de la production orale de manière asynchrone : les candidats sont enregistrés en centre d'examen lors de la passation du test, et cet enregistrement est évalué ultérieurement de manière semi-automatique. Le TOEFL dit garantir la qualité de l'évaluation en permettant aux évaluateurs humains de se concentrer sur le contenu de la production, sans être biaisés par les apparences : "*No matter who you are, or how you sound, you can be 100% confident that the only thing our test raters score is all your hard work and English skills.*" (ETS.org[4]).

La section de production orale du TOEFL iBT se compose de 4 questions qui simulent des situations de la vie réelle de l'étudiant. Elles peuvent porter sur une thématique précise, mais aucune connaissance sur le sujet n'est requise. Le temps total estimé pour la section de production orale est de 16 min. Après chaque question, le candidat dispose d'un temps de préparation de 15 à 30 s, puis doit enregistrer sa réponse au microphone pendant 45 à 60 s selon l'exercice.

- **Question 1 :** Le candidat doit se positionner par rapport à un cas présenté, en exprimant ses préférences et en argumentant son discours. L'énoncé est écrit à l'écran ; le candidat dispose de 15 s de préparation et 45 s pour donner sa réponse. Exemple d'énoncé tiré d'une vidéo tutoriel : "*Some people think it is more fun to spend time with friends in restaurants or cafés. Others think it is more fun to spend time with friends at home. Which do you think is better? Explain why.*" (ETS.org[5])
- **Questions 2 à 3 :** Elles combinent la production orale avec la compréhension de l'oral et de l'écrit :
  - **Question 2 :** Le candidat lit un court texte à propos de la vie étudiante, par exemple une annonce écrite sur un panneau d'annonce à l'université, puis il écoute une conversation entre deux personnes à propos de ce texte, où l'un des locuteurs donne son avis. Le candidat doit alors résumer l'avis de la personne en 60 s, après un temps de préparation de 30 s.
  - **Question 3 :** Le candidat lit un texte à propos d'une notion académique donnée, puis écoute un bref extrait de cours sur le même sujet, et doit ensuite expliquer la notion présentée et comment l'exemple donné dans la vidéo illustre ce concept. Temps de préparation 30 s, temps de réponse 60 s.

[4] https://www.ets.org/toefl/test-takers/ibt/scores.html, consulté le 22/07/2024
[5] https://www.ets.org/toefl/test-takers/ibt/about/content/speaking.html, consulté le 22/07/2024

- **Question 4 :** Le candidat écoute un nouvel extrait de cours et doit le résumer en listant les points mentionnés par l'enseignant. Temps de préparation 20 s, temps de réponse 60 s.

Plusieurs conseils sont donnés aux candidats : parler de manière continue pendant 45 secondes, sans se répéter et sans parler trop vite ; éviter les faux départs et les arrêts brutaux qui rendent le flux de parole saccadé ; connecter et varier ses arguments, bien noter les arguments donnés par la personne de la conversation et les mentionner dans la réponse.

La grille complète d'évaluation de la production orale du TOEFL iBT est disponible en ligne[6] et donnée en Annexe A.3. Elle est séparée en deux parties : *Independent Speaking Rubric* pour les questions 1 et 4, et *Integrated Speaking Rubric* pour les question 2 et 3. Chaque question est évaluée sur quatre critères de manière holistique sur une échelle de 0 à 4. Il y a trois critères différents : *Delivery*, pour la qualité de la prononciation et la fluidité de la parole ; *Language use*, pour la précision lexicale et grammaticale ; et *Topic development*, pour la précision et la clarté de la réponse formulée par le candidat. Concentrons-nous sur la rubrique *Delivery* (*cf.* tableau 1.1). Les descripteurs mettent en avant l'effort requis par l'auditeur pour comprendre et l'« intelligibilité » du locuteur ("*intelligibility*", terme toutefois non défini). Au niveau 2, le locuteur est intelligible mais demande des efforts à l'auditeur (articulation peu claire, intonation étrange, rythme saccadé). Au niveau 3, il commence à être un peu plus fluide (difficultés mineures de prononciation, intonation et rythme qui peuvent demander un certain effort de la part de l'auditeur mais affectent peu l'intelligibilité). Au niveau 4, le discours est généralement fluide avec des difficultés mineures qui n'affectent pas l'intelligibilité. Les descripteurs sont pratiquement mot pour mot identiques dans la partie *Integrated Speaking Rubric*.

TOEFL ITP Assessment Series

Le *TOEFL ITP Assessment Series* est un test à visée formative destiné à évaluer les compétences des étudiants pour mieux adapter l'enseignement qui leur est proposé. La production orale est évaluée par le *TOEFL ITP Speaking test*[7]. Celui-ci dure environ 15 min et est composé d'une tâche de lecture à voix haute après écoute d'un modèle, deux questions à réponse ouverte sur un sujet familier, et une question portant sur une conversation enregistrée entre deux étudiants. Dans chaque cas, l'énoncé est écrit à l'écran et lu à voix haute, une fois la lecture terminée, un chronomètre s'active pour le temps de préparation, puis un autre pour l'enregistrement. Il n'est possible de faire

[6] https://www.ets.org/pdfs/toefl/toefl-ibt-speaking-rubrics.pdf, consulté le 29/11/2024
[7] Une version démo est disponible en ligne : https://www.ets.org/toefl/itp/prepare.html

Score	Delivery
4	Generally well-paced flow (fluid expression). Speech is clear. It may include minor lapses, or minor difficulties with pronunciation or intonation patterns, which do not affect overall intelligibility.
3	Speech is generally clear, with some fluidity of expression, though minor difficulties with pronunciation, intonation, or pacing are noticeable and may require listener effort at times (though overall intelligibility is not significantly affected).
2	Speech is basically intelligible, though listener effort is needed because of unclear articulation, awkward intonation, or choppy rhythm/pace ; meaning may be obscured in places.
1	Consistent pronunciation, stress and intonation difficulties cause considerable listener effort ; delivery is choppy, fragmented, or telegraphic ; frequent pauses and hesitations.
0	Speaker makes no attempt to respond OR response is unrelated to the topic.

Tab. 1.1 : *Grille d'évaluation de la production orale du TOEFL iBT Independent Speaking Rubric, section Delivery*

une pause qu'entre les questions. Seule la conversation de la dernière question n'est pas transcrite. Comme pour le TOEFL iBT, le temps de préparation est limité (entre 30 s et 60 s selon les questions) et le temps d'enregistrement est fixé entre 45 s et 60 s.

La grille d'évaluation de la production orale pour le TOEFL ITP est accessible en ligne[8] et donnée en Annexe A.4. Elle décrit quatre niveaux de compétence de A2 à C1. Les descripteurs semblent être un mélange de ceux du TOEFL iBT et du CECRL de 2018. Au niveau A2, le locuteur est intelligible sur des sujets familiers, mais requiert un certain effort de la part de l'auditeur ; forte influence de la L1 sur la prononciation et l'accent lexical, discours entrecoupé de nombreuses pauses et faux-départs. Au niveau B1, l'accent, l'intonation et le rythme commencent à être maîtrisés mais restent parfois influencés par la L1. En B2, la parole est globalement fluide est bien rythmée *("well-paced")* malgré quelques hésitations ; l'accent et l'intonation sont maîtrisés malgré quelques erreurs. En C1, enfin, la parole est fluide, sans effort ni hésitation ; l'accent et l'intonation sont utilisés de manière stratégique.

Autant pour le TOEFL iBT que ITP, les premiers niveaux (respectivement 1 et A2) sont caractérisés par des difficultés de prononciation, d'accentuation et d'intonation dues à une forte influence de la L1, un rythme saccadé *("choppy, fragmented, telegraphic")* et de nombreuses pauses et hésitations demandant un effort important pour comprendre. Ces difficultés sont présentes dans une moindre mesure au niveau 2 ou B1, mais toujours avec une forte influence de la L1. On constate un changement clair au niveau 3 ou B2, où la parole devient globalement fluide avec une bonne maî-

[8] https://www.ets.org/pdfs/toefl/toefl-itp-speaking-descriptors.pdf, consulté le 29/11/2024

trise de l'accent, de l'intonation et du rythme, et seulement des erreurs mineures qui s'estompent encore au dernier niveau.

1.1.4 Descripteurs du TOEIC

Le *Test of English for International Communication* (TOEIC) est un test également produit par *Educational Testing Service*. Nous nous intéresserons ici à deux différentes versions : le *TOEIC Speaking Test*, qui met l'accent sur la communication en milieu professionnel, et le *TOEIC Bridge Speaking Test*, adapté pour les plus petits niveaux. Les deux se passent en ligne et en autonomie.

Le *TOEIC Speaking Test* se compose de 11 tâches et dure environ 20 min. Deux tâches sont des lectures à voix haute (45 s de préparation, 45 s d'enregistrement), deux autres sont une description d'image (45 s de préparation, 30 s d'enregistrement), les trois suivantes sont une simple question (3 s de préparation, 15 s ou 30 s de réponse), suivies de trois autres questions relatives à un court texte (45 s de lecture), et la dernière tâche demande au candidat d'exprimer son opinion (45 s de préparation, 60 s d'enregistrement).

Le TOEIC met à disposition une grille de descripteurs de niveaux[9] ainsi qu'une grille d'évaluation pour chaque type de tâche[10], données en Annexe A.5. La grille de descripteurs présente huit niveaux de compétence (de 1 à 8). On notera des difficultés constantes de prononciation, d'accentuation et d'intonation au niveau 4, une prononciation peu claire ou une intonation ou un accent inappropriés au niveau 6, ou encore de longues pauses et de fréquentes hésitations aux niveaux 4 et 5. Il est également fait mention de difficultés pour comprendre (niveaux 2 à 4) qui s'estompent peu à peu pour laisser place à une parole "*generally intelligible*" (niveau 5), "*intelligible*" (niveau 6) et "*highly intelligible*" (niveau 7).

En parallèle de ces huit descripteurs, on trouve une grille d'évaluation pour chaque type de tâche. Pour la lecture à voix haute, la performance du candidat est évaluée sur une échelle de quatre niveaux (de 0 à 3) et selon deux critères : *Pronunciation* et *Intonation and Stress*. Il est fait ici mention d'intelligibilité et du degré d'influence de la L1, de l'utilisation plus ou moins appropriée de pauses, d'emphase et de l'intonation. L'évaluation des réponses à la description d'images et aux questions est également effectuée sur quatre niveaux, mais plus orientée sur l'adéquation de la réponse en termes de contenu, de choix de vocabulaire et de structures utilisées. Il est toutefois toujours fait mention d'intelligibilité du locuteur et de l'effort de compréhension

[9] https://www.ets.org/pdfs/toeic/toeic-speaking-writing-score-descriptors.pdf, (22/11/2024)
[10] https://www.ets.org/pdfs/toeic/toeic-speaking-writing-examinee-handbook.pdf (idem)

demandé à l'auditeur. Quant à l'évaluation de l'expression d'opinion, enfin, elle est effectuée sur six niveaux plus détaillés, qui reprennent mot pour mot la grille du TOEFL iBT (*cf.* Annexe A.5).

Le *TOEIC Bridge Speaking Test* évalue les compétences communicationnelles aux niveaux débutant et intermédiaire. Les candidats répondent à des questions simples sur des sujets familiers et utilisent des phrases pour décrire des événements de la vie quotidienne. Ils peuvent être amenés à expliquer brièvement leur opinion ou leurs projets, et raconter des histoires simples. Il est mentionné dans le livret de l'examinateur[11] qu'il est attendu des candidats de pouvoir prononcer les mots de manière à être compris par un locuteur anglophone, en utilisant l'intonation, l'accent et les pauses pour rythmer *("pace")* la parole et faciliter la compréhension *("contribute to comprehensibility")*.

Le test est composé de huit tâches et dure environ 15 min. Il comprend deux tâches de lecture à voix haute, deux descriptions d'image, une tâche de type *listen and retell*, un enregistrement de message vocal, une narration d'histoire sur la base d'une suite d'images et la formulation d'une recommandation sur la base d'un texte court donné à l'écrit.

On trouve ici encore une grille d'évaluation pour chaque type de tâche, mais toutes relativement similaires. Elles proposent 4 à 5 niveaux de compétence, et se concentrent sur le degré de complétion de la tâche et la pertinence du vocabulaire et des structures utilisées. Au niveau de la prononciation, l'évaluation reste très subjective et se contente globalement de varier les adverbes : "*mostly unintelligible*" (niveaux 1), "*sometimes unintelligible*" (niveaux 2), "*generally intelligible*" (niveaux 3). L'effort demandé à l'auditeur est aussi clairement mentionné : "*requires listener effort to understand*" (niveaux 1), là encore avec différents adverbes selon les niveaux. La grille d'évaluation de la lecture à voix haute contient un peu plus d'éléments relatifs à la prononciation : "*intonation and stress are somewhat appropriate*" (niveau 1), "*mostly appropriate*" (niveau 3) ; "*lapses and/or other language influence are present*" (niveau 2), "*other-language influence does not affect overall intelligibility*" (niveau 3). De manière générale, l'évaluation porte sur l'intelligibilité du locuteur et l'effort requis pour le comprendre. Précisons qu'il est mentionné page 37 du livret de l'examinateur que "*intonation and stress refer to your ability to use emphases, pauses, and rising and falling pitch to convey meaning to a listener*".

[11] https://www.ets.org/pdfs/toeic/toeic-bridge-speaking-writing-examinee-handbook.pdf, consulté le 22/11/2024

1.1.5 Descripteurs de IELTS

L'*International English Language Testing System* (IELTS) est une certification internationale cogérée par l'université de Cambridge, le British Council et la société australienne IDP Education Limited. L'IELTS évalue les compétences du candidat sur les quatre habiletés sur un test d'une durée approximative de 2 h 45 min. Le test de production orale est une interview en face-à-face avec un examinateur durant 11 à 14 min. L'interview se compose de trois parties : une présentation personnelle du candidat, une discussion à partir d'une *task-card* présentant un sujet à aborder (monologue de 2 min puis questions-réponses), et une discussion approfondie sur ce sujet.

La version publique de la grille d'évaluation de la production orale du test IELTS est accessible en ligne[12] et donnée en Annexe A.6. Elle décrit neuf niveaux selon quatre critères : fluidité et cohérence, ressources lexicales, variété et précision grammaticale, et prononciation.

Concernant la fluidité, les descripteurs mentionnent principalement l'influence des pauses dans les niveaux 2 à 4 (niveau 2 : "*pauses lengthily before most words*", niveau 3 : "*speaks with long pauses*", niveau 4 : "*noticeable pauses*"), à partir de 5, la parole est fluide dans les contextes simples. Les répétitions, auto-corrections et hésitations sont mentionnés des niveaux 4 à 9 (niveau 4 : "*frequent repetition and self-correction*", niveau 6 : "*occasional repetition, self-correction or hesitation*", niveau 9 : "*rare repetition or self-correction ; any hesitation is content-related rather than to find words or grammar*").

Les descripteurs de prononciation sont limités et restent difficiles à interpréter. Il est fait mention de "*pronunciation features*" et "*mispronunciations*" sans plus de détails. La perception côté auditeur apparaît brièvement (niveau 4 : "*mispronunciations are frequent and cause some difficulty for the listener*", niveau 9 : "*effortless to understand*"), de même pour l'intelligibilité (niveau 8 : "*L1 accent has minimal effect on intelligibility*"). Il n'est fait nulle part mention d'accentuation ou d'intonation.

Premières conclusions

La prononciation apparaît comme un critère clé de l'évaluation de la production orale dans les principaux tests certificatifs de l'anglais. Ces tests, ainsi que le CECRL, mettent fortement l'accent sur l'intelligibilité du locuteur et l'effort demandé à l'auditeur pour le comprendre. La fluidité de la parole est un aspect qui revient souvent (CECRL, TOEFL, TOEIC, CLES), notamment l'utilisation des pauses (TOEFL, TOEIC, IELTS, CLES) ; l'accent *(stress)* est souvent mentionné (CECRL, TOEFL,

[12] https://assets.cambridgeenglish.org/webinars/ielts-speaking-band-descriptors.pdf

TOEIC), ainsi que l'intonation (CECRL, TOEIC, IELTS, CLES). Si ces paramètres sont explicités, le jugement du caractère approprié de leur utilisation est quant à lui souvent laissé à l'évaluateur (CLES B2 : « exprime ses idées avec fluidité », TOEFL iBT niveau 4 : "*generally well-paced flow*", IELTS niveau 8 : "*wide range of pronunciation features*", CECRL A2 : « les traits prosodiques (par ex. l'accent tonique) des mots familiers et quotidiens et des énoncés simples sont convenables », CECRL B2 : « peut généralement placer correctement l'accent »). Lorsque des précisions sont données pour aider l'évaluateur, elles restent relativement limitées. Par exemple, le caractère fluide de la parole est déterminé par l'absence de longues pauses pour le CLES, ou de répétitions, auto-corrections et hésitations non-liées au contenu pour IELTS. De manière générale, même lorsqu'il est fait mention d'une influence notable de la L1, ce sont avant tout les phénomènes affectant l'intelligibilité qui sont considérés comme problématiques.

À l'issue de ce tour d'horizon, nous avons donc une idée plus précise des paramètres à cibler lors de l'évaluation de la prononciation : tout ce qui affecte l'intelligibilité du locuteur, et en premier lieu la fluidité (relative à la présence de pauses et d'hésitations), et les traits prosodiques comme l'accentuation et l'intonation. Qu'en est-il maintenant des systèmes d'évaluation automatique de la prononciation ? Ciblent-ils les mêmes paramètres ? et sur quelles bases reposent leurs jugements ?

1.2 Évaluation automatique

Les premiers systèmes d'évaluation automatique de la prononciation sont arrivés dans les années 90 avec les débuts de la reconnaissance automatique de la parole. Le système Autograder (Bernstein et al., 1990) est pionnier dans le domaine : il présente une liste de questions à choix multiple, pour lesquelles l'apprenant est amené à lire à voix haute l'une des options de réponse. La machine identifie alors l'option qui a été prononcée, et donne un score basé sur le nombre de mots correctement reconnus. Un peu plus tard, les systèmes VILTS (*Voice Interactive Language Training Systems*, Neumeyer et al., 1996, Franco et al., 1997) permettent de donner n'importe quel texte à la machine pour le faire lire à l'apprenant. Les scores sont calculés à partir de mesures segmentales (reconnaissance des phonèmes) et suprasegmentales (durée des phonèmes, des syllabes ou débit de parole). Les premiers systèmes qui évaluent la parole spontanée arrivent dans les années 2000, et se focalisent sur la fluence de la parole (Cucchiarini et al., 2002), mais la parole spontanée reste toutefois marginale par rapport à la parole lue.

Après un fort engouement pour l'évaluation automatique de la prononciation à la fin des années 90, la discipline s'essouffle à cause d'une mauvaise fiabilité des systèmes

alors commercialisés (Witt, 2012). Elle revient toutefois rapidement à la charge avec la généralisation des smartphones et l'amélioration des systèmes de reconnaissance de parole à la fin des années 2000. Notons la création de la conférence *Speech and Language Technology for Education* (SLaTE) au sein de l'*International Speech Communication Association* (ISCA) en 2007, qui se consacre spécifiquement à l'apprentissage des langues et les technologies de la parole (Ellis & Bogart, 2007).

Vingt ans après ces débuts, à quoi ressemblent les systèmes et comment évaluent-ils la prononciation ?

Commençons par distinguer deux types d'évaluation. L'évaluation certificative d'une part *(high-stake assessment)*, dont l'objectif premier est de déterminer le niveau du candidat et peut être la condition d'obtention d'un diplôme, d'un emploi, voire d'un visa ; et l'évaluation formative d'autre part *(low-stake assessment)*, qui a pour but d'identifier les difficultés de l'apprenant et lui fournir un feedback pour l'aider à progresser. Il apparaît dans la littérature que ces deux types d'évaluation se distinguent par les techniques qu'elles mettent en œuvre pour évaluer la prononciation : la première choisit généralement d'entraîner des modèles à prédire le score global du locuteur sur la base d'évaluations humaines, tandis que la deuxième cherche plutôt à identifier et mesurer des phénomènes cibles afin de proposer un feedback formatif à l'utilisateur.

1.2.1 Évaluation certificative

Ces dernières années, de plus en plus de tests certificatifs se sont équipés de systèmes d'évaluation automatique de la production orale. Ces outils sont conçus pour prédire le score d'un candidat à partir d'un enregistrement audio, on parle de *machine scoring* (Davis & Papageorgiou, 2021). Un grand nombre de productions d'apprenants est évalué par des experts sur des critères similaires à ceux mentionnés dans la section précédente, et un modèle est entraîné à prédire les scores donnés par les évaluateurs à partir de mesures automatiques diverses. Educational Testing Service fait partie des leaders mondiaux en la matière avec des entraînements sur plusieurs centaines de milliers d'enregistrements de candidats (Loukina & Yoon, 2019). Une fois le modèle entraîné, le système est capable de prédire le score d'un nouvel enregistrement qui n'a pas été évalué manuellement.

Les mesures utilisées sont avant tout des paramètres en lien avec la fluence, mais de plus en plus souvent combinés avec des paramètres spectraux (Evanini & Wang, 2013 ; Fontan et al., 2018), lexicaux (Yoon et al., 2012) ou syntaxiques (Bhat & Yoon, 2015 ; L. Chen & Zechner, 2011 ; Loukina et al., 2015). Parmi les paramètres de fluence utilisés, on retrouve généralement le débit de parole et d'articulation, la fréquence de pauses pleines et silencieuses et leur durée moyenne, le nombre de syllabes par unité

rythmique, la durée des syllabes ou des voyelles, ou encore les variations d'intonation et d'intensité. Ces techniques combinent souvent de nombreux paramètres (77 pour Coutinho et al., 2016, 75 pour Loukina et al., 2015), mais ce sont souvent les mêmes qui obtiennent la meilleure corrélation avec les jugements humains : le débit de parole et d'articulation, la proportion de pauses et leur durée moyenne, la longueur des segments entre pauses.

Dès les premiers systèmes de ce type dans les années 2000, la corrélation humain/machine est comparable à la corrélation inter-évaluateurs : 0,8 pour Neumeyer et al. (2000) et Cucchiarini et al. (2002), 0,7 pour Moustroufas et Digalakis (2007). On tourne autour des mêmes valeurs aujourd'hui, même en parole spontanée : 0,8 pour Fu et al. (2020), 0,8-0,9 pour Shen et al. (2021), 0,8 encore pour Saito et al. (2022). Si la corrélation avec les évaluations humaines est élevée, ces outils restent toutefois limités lorsqu'il s'agit d'évaluer des paramètres de plus hauts niveaux, comme l'organisation du discours, la précision grammaticale ou la cohérence de l'énoncé (Isaacs, 2018). Aussi sont-ils souvent combinés avec des jugements humains pour garantir une meilleure fiabilité des résultats, tout en bénéficiant des avantages de l'évaluation automatique – on parle de *hybrid human-machine scoring*. On considère généralement trois approches : l'approche hybride confirmatoire, l'approche contributive parallèle et l'approche contributive divergente (Davis & Papageorgiou, 2021). Dans la première, l'évaluation automatique est seulement utilisée pour confirmer l'évaluation manuelle d'un examinateur ; si l'écart entre les deux évaluations est jugé trop grand, un second examinateur est mobilisé. Dans l'approche contributive parallèle, deux scores holistiques, un manuel et un automatique, sont combinés pour déterminer le score final. Dans l'approche contributive divergente, enfin, les évaluations humaines et automatiques portent sur des aspects différents et complémentaires de la production, typiquement de bas niveau pour la machine et de plus haut niveau pour l'humain.

Parmi les tests certificatifs qui intègrent des systèmes de prédiction de scores pour l'évaluation de la production de l'oral en anglais, on trouve le TOEFL iBT avec son système SpeechRater, intégré selon l'approche contributive parallèle (Evanini & Zechner, 2019) ; les Versant English Tests (complètement automatisés, Pearson Education, 2022), le Pearson Test of English (Pearson PTE, 2024), le Duolingo English Test (complètement automatisé également, Cardwell et al., 2024), le Cambridge Assessment English Linguaskill General Speaking Test avec son système Custom Automated Speech Engine (Xu et al., 2021). Le bénéfice de l'utilisation de tels systèmes pour les organismes certificatifs est grand : s'il est coûteux à concevoir, il est vite rentabilisé par les économies en termes de ressources humaines nécessaires pour évaluer les productions des candidats, et permet de réduire drastiquement le temps de délivrance des résultats qui se compte en semaines pour les évaluations humaines (Isaacs, 2018). Par ailleurs, la systématicité de l'évaluation est souvent mise en avant comme garante d'une évaluation plus équitable.

Si ces systèmes se révèlent performants pour prédire un score global à partir d'une production orale, ils sont toutefois peu exploitables en contexte diagnostique, étant donné que les paramètres sur lesquels ils se basent sont majoritairement de bas niveau. Si le score du candidat est influencé par le débit de parole ou la fréquence des pauses, ce n'est pas pour autant qu'il doit parler plus vite ou faire moins de pauses. Ces phénomènes sont une conséquence de difficultés en amont, mais pas nécessairement un problème en soi. En contexte formatif, les systèmes d'évaluation ont donc dû adopter d'autres approches pour pouvoir donner un feedback pédagogique à l'utilisateur.

1.2.2 Évaluation formative

Un grand nombre d'applications d'apprentissage des langues proposent aujourd'hui des fonctionnalités d'évaluation de la prononciation. Les plus en vogue en 2022 étaient Duolingo[13], Memrise[14], Babbel[15], Busuu[16] ou Rosetta Stone[17]. D'autres applications sont dédiées à la prononciation de l'anglais, comme ELSA[18] ou IELTS Speaking Practice[19] par exemple. Cette sous-section présente la manière dont ces applications évaluent la prononciation, le type d'activités qu'elles proposent et les feedbacks qu'elles donnent aux utilisateurs. Nous nous basons ici sur une description plus complète des applications proposée dans un chapitre d'ouvrage (Coulange, 2023).

En 2022, l'activité de production orale la plus courante dans les applications d'apprentissage des langues consistait à lire à voix haute ou à répéter un mot ou un énoncé en appuyant sur un bouton d'enregistrement. La production est analysée en temps réel et un feedback immédiat est affiché. Toutes les applications mentionnées dans le paragraphe précédent proposent ce type d'activité, accompagnée d'un modèle audio de ce qui doit être prononcé et de la transcription orthographique affichée par défaut, sauf pour Rosetta Stone qui utilise parfois des images sans texte ni audio pour éliciter la parole. Certaines applications affichent également une traduction dans la langue de l'utilisateur, par défaut, comme Memrise ou Babbel, ou sur demande et mot à mot comme Duolingo. L'audio peut être accompagné d'une vidéo ou d'une image fournissant des indices contextuels. ELSA propose également une transcription phonétique, une fonctionnalité qui semble absente des autres applications. ELSA et Memrise permettent aussi de ralentir la vitesse de lecture.

[13] DuoLingo Inc. (2022). https://www.duolingo.com/
[14] Memrise (2022). https://www.memrise.com/
[15] Babbel (2022). https://uk.babbel.com/
[16] Busuu Online S.L. (2022). https://www.busuu.com/
[17] Rosetta Stone Inc. (2022, v5.0.37). https://www.rosettastone.com/
[18] Elsa Speak (2022). https://elsaspeak.com/
[19] SpeechAce LLC (2022). https://www.speechace.com/

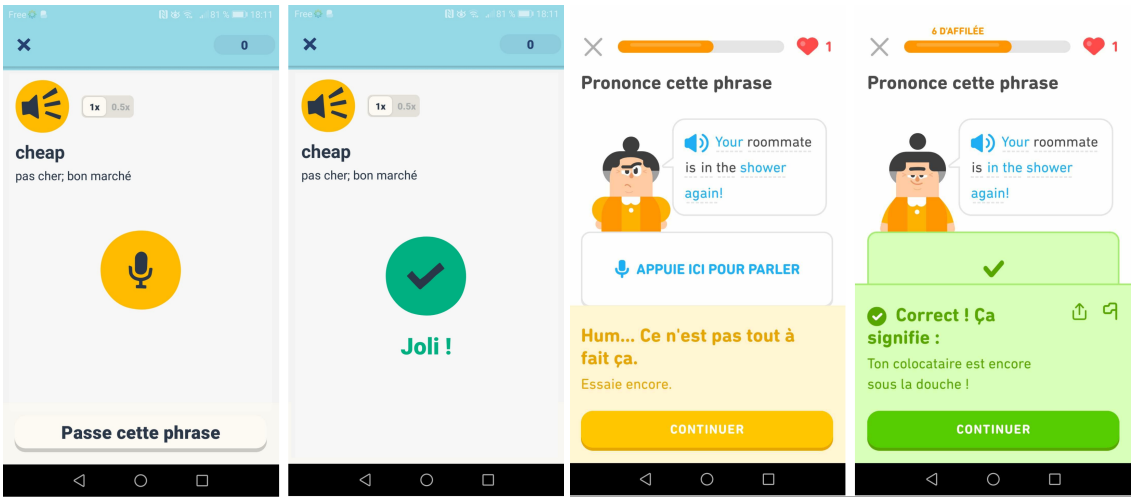
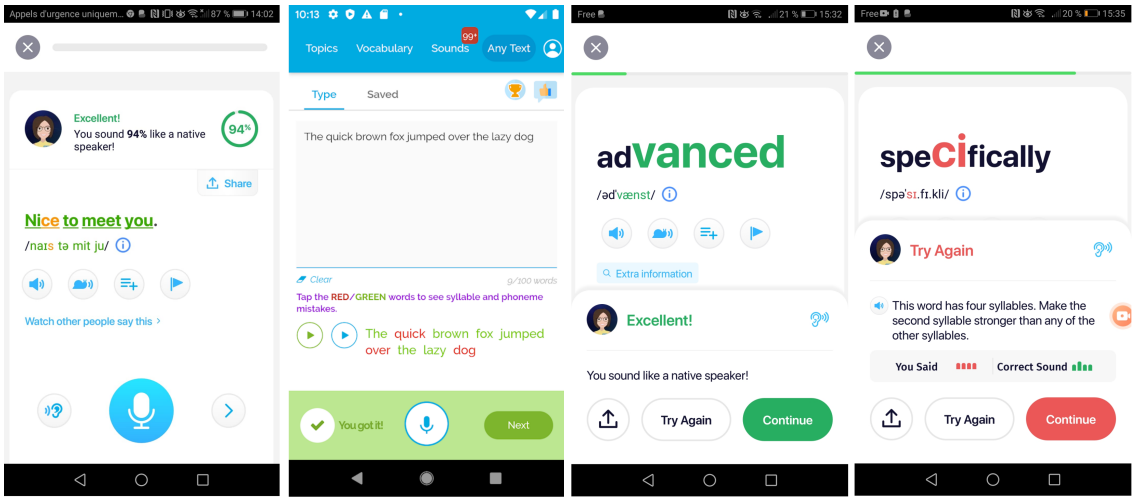


FIG. 1.1 : Captures d'écran de Memrise (gauche) et Duolingo (droite) en novembre 2022.

Dans la majorité des cas, les feedbacks donnés à l'utilisateur sont binaires (correct ou incorrect). Dans Duolingo, par exemple, un écran vert s'affiche avec des félicitations lorsque la production est validée, et dans le cas contraire un écran orange indique à l'apprenant qu'il peut mieux faire, sans toutefois donner de conseils pour améliorer sa prononciation (*cf.* figure 1.1). D'autres applications affichent un pourcentage de réussite indiquant dans quelle mesure les mots ont été reconnus par le système. ELSA et IELTS Speaking Practice vont un peu plus loin en affichant les mots ou les lettres en couleurs : vert pour les mots ou phonèmes reconnus correctement, orange quand ce n'est pas tout à fait bon, et rouge quand le mot ou le phonème est incorrect ou manquant (*cf.* figure 1.2a). En cliquant sur un mot, l'apprenant peut voir les phonèmes attendus et ceux qui ont été identifiés par le système, accompagnés de conseils explicites pour prononcer chaque phonèmes. ELSA propose également une activité dédiée à l'accent lexical. Un mot isolé apparaît à l'écran avec la syllabe à accentuer écrite en gros caractères. Après l'enregistrement, cette syllabe est colorée en vert ou en rouge selon la syllabe accentuée par l'utilisateur. Une représentation visuelle des syllabes est également affichée sous forme de barres plus ou moins hautes pour symboliser la position de l'accent primaire (*cf.* figure 1.2b).

En septembre 2022, ELSA a déployé une nouvelle fonctionnalité premium appelée Speech Analyzer, permettant aux étudiants d'enregistrer une production orale libre et d'obtenir un score global de production orale, ainsi que des scores détaillés pour la prononciation, l'intonation, la fluidité, la grammaire et le vocabulaire. ELSA fournit également des prédictions de scores pour IELTS, TOEFL, Pearson ainsi que le niveau CECRL. Comme pour la lecture à voix haute, Speech Analyzer identifie les erreurs segmentales et propose des conseils pour les corriger. L'outil calcule égale-



(a) ELSA (gauche) et IELTS Sp. Pr. (droite) *(b) Accent lexical sur ELSA*

Fig. 1.2 : *Captures d'écran de novembre 2022*

ment des scores d'intonation, de débit de parole et de pauses, exprimés sous forme de pourcentages (*cf.* figure 1.3).

À l'exception de l'exercice de détection de l'accent lexical d'ELSA, tous les systèmes mentionnés ci-dessus calculent un score de type *Goodness Of Pronunciation* (GOP) (Witt, 1999), basé sur le niveau de confiance d'un système de reconnaissance vocale. La réponse est considérée comme correcte lorsque le score de confiance dépasse un certain seuil. Dans IELTS Speaking Practice ou ELSA, la reconnaissance phonémique non contrainte permet à l'apprenant de voir quels phonèmes ont été reconnus par le système, bien qu'il soit limité aux phonèmes de l'anglais et au nombre de phonèmes du mot ou de l'énoncé de référence. Aucune information n'a toutefois été trouvée sur le fonctionnement des outils de détection de la syllabe accentuée pour ELSA.

Du côté de la recherche publique, on trouve un grand nombre d'études proposant des systèmes pour évaluer la prononciation. Elles aussi proposent dans leur écrasante majorité le calcul d'un score de type GOP, sur la base de la reconnaissance des phonèmes. Ces scores se déclinent toutefois dans une grande diversité : certaines études proposent d'adapter le système de reconnaissance à la parole L2, en augmentant un lexique phonétisé avec des erreurs phonologiques typiques (*Extended Recognition Network*, Bada et al., 2020 ; Lee et Glass, 2015), ou en adaptant les modèles acoustiques à partir d'enregistrements de la langue maternelle des apprenants (Goronzy et al., 2004 ; Tan, 2008), ou directement avec de la parole L2 (Duan et al., 2017 ; W. Li et al., 2016). D'autres encore proposent de comparer la reconnaissance d'un système entraîné sur de la parole native avec celle d'un système adapté pour la parole L2, le score est alors

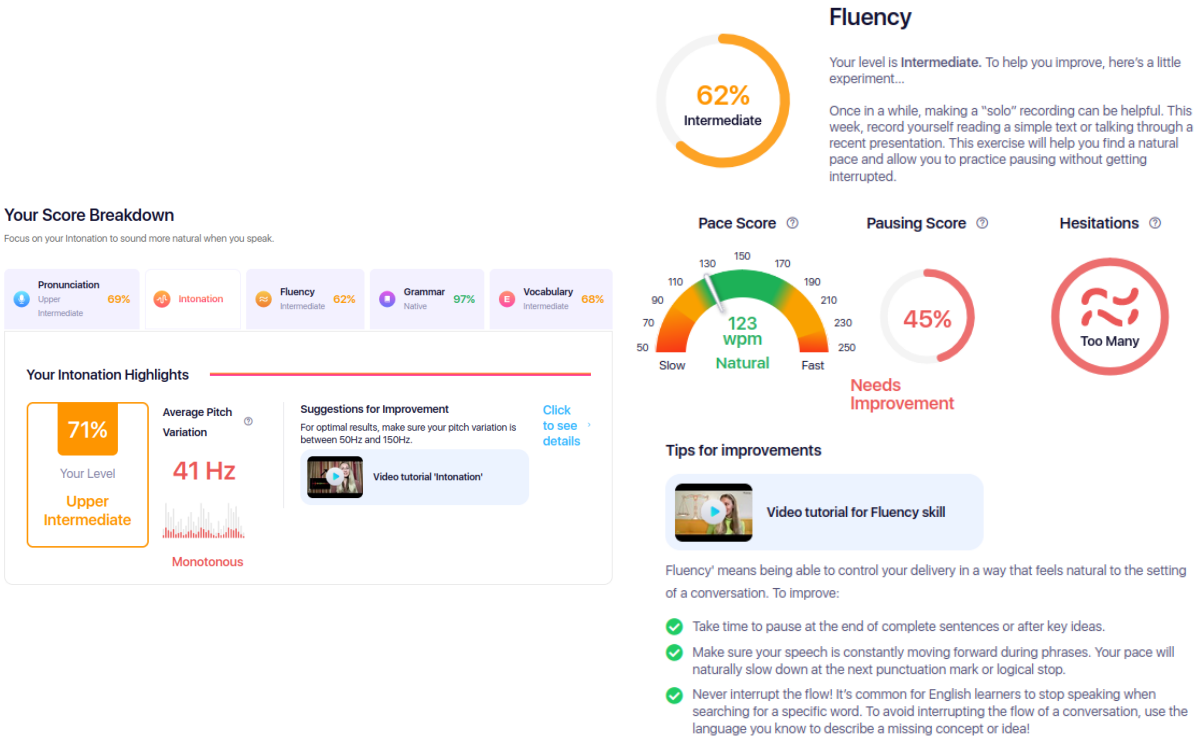


FIG. 1.3 : Captures d'écran de Speech Analyzer (ELSA) en novembre 2022.

basé sur la différence des deux sorties et permet de se passer du texte de référence (*Reference free Error Rate*, Fu et al., 2020 ; Naijo et al., 2021). D'autres systèmes s'emploient à comparer des mesures acoustiques (durées, débit, intonation, traits phonétiques etc.) directement avec un modèle, qu'il s'agisse d'un enregistrement du même énoncé par un locuteur natif (Arias et al., 2010 ; Ding et al., 2020), ou un modèle appris sur un ensemble d'enregistrements pour permettre plus de variabilité (Truong et al., 2018 ; Wang et al., 2015). À part Fu et al. (2020), toutes les études mentionnées ici se concentrent sur l'évaluation de mots ou de phrases lues.

Dans la majorité des systèmes d'évaluation formative, la prononciation est mesurée en termes de distance par rapport à un modèle natif. ELSA va même jusqu'à afficher un pourcentage dans ses feedbacks : "*You sound 94% like a native speaker!*" (*cf.* figure 1.2a). Ces systèmes reposent souvent sur la reconnaissance automatique de la parole pour générer leurs scores, qu'il s'agisse de taux de reconnaissance de mots ou de phonèmes. Certains intègrent des dimensions prosodiques, telles que le débit de parole ou la fréquence des pauses, mais les innovations dans ce domaine restent limitées.

Or, comme le souligne Isaacs (2018), "*the element of accent reduction that the software is targeting may be incompatible with helping learners become intelligible*" (p. 20). En effet, non seulement l'objectif de « parler comme un natif » est ambitieux et souvent irréaliste, mais il n'améliore pas nécessairement l'intelligibilité d'un locuteur (Derwing & Munro, 2015). Certains éléments évalués par ces systèmes automatiques sont secondaires pour la communication. Il semble donc plus pertinent d'encourager les apprenants à se concentrer sur les phénomènes qui entravent ou perturbent, voire rendent impossible, la compréhension mutuelle (Isaacs, 2018).

Conclusion

Les exigences des tests certificatifs accordent une place centrale à l'intelligibilité et la compréhensibilité du locuteur. Le rythme, la fluidité, les pauses, l'intonation ou l'accent sont régulièrement cités comme éléments clés pour estimer le niveau de compétence du locuteur.

Les systèmes d'évaluation automatique adoptés par ces tests offrent des solutions techniques pour estimer un niveau de production orale. En s'appuyant sur des mesures de bas niveau, comme le débit de parole, la fréquence des pauses ou le nombre de mots, ces systèmes parviennent à prédire des scores globaux avec une précision comparable à celle des évaluateurs humains. Cependant, leur incapacité à juger des compétences de haut niveau, comme la cohérence ou la précision de la réponse, limite leur usage à une utilisation complémentaire à l'évaluation humaine. Par ailleurs, les mesures effectuées ne sont pas directement exploitables dans un contexte formatif car elles ne ciblent pas spécifiquement les phénomènes qui perturbent la communication.

Dans le cadre de l'évaluation formative, les systèmes automatiques se concentrent plutôt sur des scores basés sur la reconnaissance automatique de la parole, de type *Goodness of Pronunciation*, qui mesure une proximité à un modèle natif. L'évaluation porte majoritairement sur de la parole lue ou répétée, délaissant la parole spontanée, pourtant essentielle au développement de la compétence communicative.

Un décalage important persiste entre les objectifs pédagogiques actuels et les approches adoptées par les systèmes automatiques. Si l'on attend du locuteur qu'il soit intelligible et facilement compris par l'auditeur, les systèmes d'évaluation formative restent encore beaucoup centrés sur la comparaison avec un modèle natif. Cette divergence souligne la nécessité de proposer des outils formatifs mieux alignés sur les objectifs pédagogiques, afin qu'ils puissent apporter un complément efficace aux enseignements classiques.