Synergies France n° 17 - 2023 p. 135-152

GERFLINT
ISSN 1766-3059
ISSN en ligne 2260-7846

Conception d'un module de diagnostic automatique de la prononciation

Sylvain Coulange
Université Grenoble Alpes, France
sylvain.coulange@univ-grenoble-alpes.fr
https://orcid.org/0000-0002-9728-1181

Reçu le 15-07-2022 / Évalué le 24-10-2022 / Accepté le 24-02-2023

Résumé

Cet article présente les premières pistes de conception d'un module de diagnostic automatique de la prononciation en anglais langue étrangère, dans le cadre du Système d'Évaluation en Langues à visée Formative (SELF), porté par l'Université Grenoble Alpes. L'état de l'art des systèmes d'évaluation automatique de la prononciation fait ressortir un écart entre les phénomènes mesurés et les objectifs pédagogiques en termes de prononciation. Notre système se donne pour objectif d'identifier et mesurer les phénomènes de prononciation jugés comme impactant la compréhensibilité du locuteur, en se focalisant dans ce travail sur la réalisation de l'accent lexical, ainsi que le positionnement des pauses dans le flux de parole. Le calibrage du système se fera grâce au corpus d'interactions orales du test certificatif CLES entre les niveaux B1 et B2. Cet article présente le constat de départ, le corpus et les premiers traitements réalisés.

Mots-clés : évaluation diagnostique, prononciation, compréhensibilité, évaluation automatique

Creating an automatic pronunciation diagnostic module

Abstract

This article presents the first steps in the design of an automatic pronunciation diagnostic module for English as a foreign language, within the framework of the Système d'Évaluation en Langues à visée Formative (SELF), headed by the Université Grenoble Alpes. The state-of-the-art automatic pronunciation evaluation systems shows a gap between the measured phenomena and the pedagogical objectives in terms of pronunciation. Our system aims to identify and measure the pronunciation phenomena supposedly impacting the comprehensibility of the speaker, while focusing on the realization of the lexical stress as well as the positioning of pauses in the speech flow. The calibration of the system will be done thanks to the corpus of oral interactions of the CLES certification test between levels B1 and B2. This article presents the initial findings, the corpus and the first analysis performed.

Keywords: diagnostic assessment, pronunciation, comprehensibility, automatic assessment

Introduction

Le Système d'Évaluation en Langues à visée Formative (désormais SELF) est un test de positionnement en ligne développé à l'Université Grenoble Alpes dans le cadre du projet IDEFI Innovalangues (ANR-11-IDFI-0024, 2012-2020). Le SELF est actuellement déployé en 6 langues (anglais, italien, espagnol, français, japonais et mandarin) et utilisé dans une vingtaine d'universités françaises ainsi que quelques institutions à l'international. L'évaluation porte sur trois habiletés langagières : la compréhension de l'oral, la compréhension de l'écrit et l'expression écrite courte.

En parallèle des tests de positionnement, le SELF souhaite proposer des modules diagnostiques capables de fournir une rétroaction détaillée sur les acquis et les difficultés des apprenants dans une compétence donnée. Le premier module envisagé est un dispositif d'aide à l'évaluation formative de la production orale, exploitant les technologies de traitement automatique de la parole pour identifier et mesurer les phénomènes impactant la compréhensibilité du locuteur. Ce chantier est initié par un travail de thèse de doctorat[1] au sein du laboratoire de Linguistique et Didactique des Langues Étrangères et Maternelles (LIDILEM) et du Laboratoire d'Informatique de Grenoble (LIG). Cet article présente les premières réflexions de ce travail.

L'évaluation de la production orale touche de nombreux aspects de la langue, de la cohérence syntaxique à la précision du vocabulaire, en passant par la pragmatique ou la cohésion du discours. S'il y a pourtant un aspect qui influence globalement l'évaluation, c'est la qualité de prononciation (Pennington, 1999). Par prononciation, nous entendrons ici tous les aspects de la parole relatifs à la phonétique et à la prosodie.

L'évaluation de la prononciation, lorsqu'elle est explicite dans les grilles d'évaluation, se résume souvent à un score global établi par l'enseignant sur la précision phonétique ou la fluence de parole (Piccardo, 2016). Toutefois, les enseignants ont rarement les outils et la formation nécessaires pour évaluer ces aspects de manière précise et systématique, et il en résulte souvent un jugement basé sur l'intuition, avec une certaine variabilité d'un évaluateur à l'autre et souvent chez un même évaluateur. Gilquin et al. (2022) notent par exemple que les enseignants non natifs ont tendance à être plus sévères que les enseignants natifs lorsqu'ils évaluent la prononciation. Par ailleurs, un évaluateur ne sera pas sensible de la même manière à la prononciation des apprenants selon son degré de familiarité avec leur accent (Didelot et al., 2019). De manière générale, un manque de formation des enseignants pour enseigner et évaluer la prononciation est clairement ressenti (Piccardo,

2016 ; Derwing, Munro, 2015 ; Baker, 2011 ; Burgess, Spencer, 2004 ; Breitkreutz et al., 2001), et il est apparu nécessaire de développer des critères d'évaluation précis, de généraliser leur utilisation et de créer des outils d'apprentissage et d'évaluation basés sur ces derniers.

En 2007, Levis constatait que les études sur l'évaluation automatique de la prononciation arrivaient rarement au stade de production pour plusieurs raisons. Pédagogiques, d'une part, étant donné l'écart entre ce que proposaient les outils de l'époque et les objectifs pédagogiques. Technologiques, d'autre part, constatant que les outils n'étaient pas capables de fournir un feedback précis et pertinent sur la prononciation. Enfin, l'auteur constatait lui aussi un manque de formation des enseignants, au niveau phonétique comme au niveau technologique (Levis, 2007). Qu'en est-il 15 ans plus tard ?

Notre travail a commencé par une exploration des systèmes d'évaluation automatique de la prononciation, ainsi qu'une réflexion sur l'évaluation de la prononciation et les critères d'évaluation d'aujourd'hui. C'est sur cette base que nous avons choisi de nous concentrer sur deux aspects de la prononciation : l'accent lexical et les pauses, la section 3 présente les paramètres acoustiques que nous avons décidé de cibler, et la section 4 la méthodologie envisagée pour mettre au point le dispositif d'évaluation.

1. Bref état de l'art des systèmes existants

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 (désormais ASR). 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 quelle option 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 et débit de parole). Les premiers systèmes évaluant la parole spontanée sont arrivés dans les années 2000, en se focalisant sur la fluence de la parole (Cucchiarini et al., 2002), mais le spontané reste très marginal encore aujourd'hui par rapport à la parole lue, bien plus prédictible.

Après un fort intérêt pour l'évaluation automatique de la prononciation à la fin des années 90, la discipline s'est essoufflée à cause d'une mauvaise fiabilité des systèmes alors commercialisés (Witt, 2012). Elle est toutefois rapidement revenue à la charge avec la généralisation des smartphones à la fin des années 2000, l'augmentation de la puissance de calcul et l'amélioration des systèmes d'ASR. Notons la création du groupe SLaTE *(Speech and Language Technology for Education)* au sein de l'ISCA *(International Speech Communication)* en 2007, qui travaille spécifiquement sur l'apprentissage des langues et les technologies de la parole (Ellis, Bogart, 2007). Pour une revue détaillée des applications antérieures à 2011, le lecteur peut consulter (Witt, 1999, O'Brien, 2006, Levis, 2007, Eskenazi, 2009 ou encore Delmonte, 2011).

Mais à quoi ressemblent les systèmes d'aujourd'hui et comment évaluent-ils la prononciation ?

Commençons par distinguer deux types d'évaluation. L'évaluation formative d'une part *(low stake assessment)*, qui a pour but d'accompagner l'apprenant dans son parcours et cibler ses acquis et ses difficultés ; et l'évaluation certificative d'autre part *(high stake assessment)*, dont les enjeux sont plus importants, puisqu'ils peuvent être la condition d'obtention d'un diplôme, d'un emploi, voire d'un visa.

Commençons par l'évaluation formative. La majorité des applications récentes d'apprentissage des langues proposent des fonctionnalités d'évaluation de la prononciation. Citons seulement les plus connues : Duolingo, Memrise, Babbel, Busuu, Rosetta Stone, LingoChamp (ex. Liulishuo). Elles fonctionnent toutes sur le même principe, à savoir la lecture d'un mot ou d'une phrase à voix haute, et fournissent un feedback binaire (type bon/mauvais, Memrise, Duolingo, Busuu) ou un score de confiance de reconnaissance (en pourcentage, LingoChamp, Rosetta Stone). Elles proposent toutes d'entendre le mot ou l'énoncé en parallèle de la lecture, d'office (Memrise, Busuu, Babbel) ou à la demande (Duolingo, Rosetta Stone). D'autres applications sont dédiées à la prononciation, comme ELSA Speak ou IELTS Speaking Practice par exemple. Il s'agit toujours de lecture et/ou de répétition, mais les rétroactions proposées sont plus détaillées : les mots ou les phonèmes sont colorés en fonction de la confiance de l'ASR : le système indique quels phonèmes ont été reconnus (parmi les phonèmes de l'anglais uniquement). Enfin, des scores spécifiques à telle ou telle catégorie de phonèmes sont proposés (ex. voyelles réduites, consonnes occlusives, phonèmes en fin de mot sur ELSA) ou des scores globaux (nombre moyen de mots correct par minute sur IELTS Speaking Practice).

On constate que la majorité de ces applications évaluent la parole lue (ou répétée) et proposent donc des protocoles d'élicitation très contraints. Quant aux rétroactions fournies à l'apprenant, si elles ne sont pas seulement binaires, elles se concentrent généralement sur les aspects segmentaux de la parole, sur la base de la reconnaissance de chaque mot ou phonème attendu, et sont présentées à l'utilisateur sans filtrage ni hiérarchisation. On peut alors se demander à juste titre si ces systèmes ont vraiment un impact positif sur la compétence des apprenants. Plusieurs études proposent une évaluation de ces outils, mais la plupart sont financées par l'entreprise qui développe l'application en question et sont bien peu critiques, peu sont publiées dans des revues avec comités de lecture et le financement ou l'appartenance des auteurs ne sont pas toujours précisés. On notera tout de même quelques rares études moins enthousiastes quant à l'efficacité de ces outils, comme Becker et Edalatishams, 2018, Krashen, 2013, 2014, ou encore Nielson, 2011, mais celles-ci sont peu nombreuses.

Du côté de la recherche pourtant, l'évaluation automatique de la prononciation est toujours assez tendance, et un grand nombre d'études sont publiées chaque année dans les revues phare comme InterSpeech ou LREC.

La majorité des études publiées depuis les années 90, et encore aujourd'hui en 2022, présente des systèmes qui basent leurs scores sur une comparaison des mots ou des phonèmes reconnus par un système d'ASR avec un texte cible. Ce type de scores est appelé *Goodness of pronunciation* et a été introduit par Witt et Young (2000). La reconnaissance est plus ou moins contrainte et plus ou moins adaptée à la parole non standard de l'apprenant, et se décline dans une formidable diversité. Il peut s'agir d'un ASR classique et d'un taux de reconnaissance du texte cible (c'est le cas des applications commerciales présentées plus haut) ; certaines études proposent d'adapter le système de reconnaissance à la parole non native, en augmentant son lexique phonétisé avec des erreurs phonologiques types (*Extended Recognition Network*, Bada et al., 2020, Lee, Glass 2015), ou en adaptant les modèles acoustiques à partir d'enregistrements de la langue maternelle des apprenants (Tan, 2008, Goronzy et al., 2004), ou directement avec de la parole L2 (Duan et al. 2017, Li et al., 2016). D'autres encore proposent de comparer la reconnaissance d'un ASR standard avec celle d'un ASR entraîné sur la parole L2, le score est alors 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 (Ding et al., 2020, Arias et al., 2010), ou un modèle appris sur un ensemble d'enregistrements pour permettre plus de variabilité (Truong et al., 2018, Wang et al. 2015).

On constate que dans la grande majorité de ces systèmes, l'évaluation de la prononciation se traduit sous la forme d'une distance par rapport à un modèle natif. Une grande partie d'entre eux se fonde sur la reconnaissance de la parole pour établir le score, qu'il s'agisse d'un taux de reconnaissance de mots ou de phonèmes, ou bien d'un score de proximité au modèle, plus difficile à interpréter. Certains combinent ces valeurs avec des mesures de débit de parole ou de fréquence des pauses pour ajouter un aspect prosodique à l'évaluation, mais peu d'innovation a été faite sur ce plan ces 20 dernières années. Nous rejoignons le constat de Detey et ses collègues (2016) quant au fait que la prosodie n'est pas suffisamment prise en compte dans l'évaluation. Rappelons enfin que presque tous les outils évaluent de la parole lue, souvent hors contexte, qu'il s'agisse d'applications commercialisées ou d'études plus exploratoires. Il s'agit même encore parfois de lecture de mots isolés (Kato et al., 2019, Truong et al. 2018). Si le défi technique d'évaluer la parole spontanée est clair, c'est davantage une sous-estimation de son importance dans la compétence en langue qui ressort de notre état de l'art. Nous faisons le même constat pour la prosodie.

Du côté certificatif, le problème est pris différemment. Plutôt que de comparer une production d'apprenant à un modèle natif, ce sont des modèles statistiques qui sont entrainés à prédire le score d'un évaluateur humain. Ces modèles de prédiction sont appris sur la base d'un grand nombre de productions d'apprenants évaluées par des experts sur différents critères (souvent globaux, du type fluence, degré d'accent, intelligibilité ou compréhensibilité), ainsi que sur des mesures automatiques diverses (avant tout des paramètres suprasegmentaux, mais de plus en plus souvent combinés avec des paramètres segmentaux (Fontan et al., 2018, Evanini, Wang, 2013), lexicaux (Yoon et al., 2012) ou syntaxiques (Loukina et al., 2015, Bhat, Yoon 2015, Chen, Zechner, 2011). Parmi les paramètres suprasegmentaux 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 un très grand nombre de paramètres (77 pour Coutinho et al., 2016, 75 pour Loukina et al., 2015), mais ce sont souvent les mêmes qui sont les plus significatifs : le débit de parole et d'articulation, la proportion de pauses et leur durée moyenne. Une fois le modèle entrainé sur le corpus, le système est capable de prédire le score d'un nouvel enregistrement qui n'a pas été évalué manuellement. 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 avec de la parole spontanée : 0,823 pour Saito et al., 2022, 0,8-0,9 pour Shen et al., 2021, 0,799 pour Fu et al., 2020. Ce sont ces systèmes qui sont généralement exploités dans les tests certificatifs automatisés, comme le Versant English Test (Pearson Education, Inc., 2022), ou le TOEFL iBT (Zechner et al., 2019).

Si ces systèmes se révèlent performants pour prédire un niveau global sur une production donnée, ils sont toutefois difficilement exploitables en contexte diagnostique. Il est en effet peu pertinent d'indiquer à l'apprenant qu'il parle trop lentement ou trop vite, ou bien qu'il fait trop de pauses, et que c'est à cause de cela que son niveau est plus faible. Ces phénomènes sont une conséquence de difficultés en amont, mais pas nécessairement un problème en soi. Ce qui paraît plus pertinent en revanche, c'est d'identifier les éléments les plus susceptibles de dégrader la compréhensibilité en situation d'interaction orale, de les localiser et de les mesurer.

2. Prise de recul sur l'évaluation de la prononciation

Si l'apprentissage de la prononciation a longtemps eu pour objectif de se rapprocher autant que possible d'un modèle natif, les récentes études s'accordent aujourd'hui sur l'importance de *se faire comprendre* et d'être capable de communiquer clairement, sans plus nécessairement devoir imiter l'accent d'un autre (Walker et al., 2021 ; Conseil de l'Europe, 2020 ; Isaacs et al., 2017). La compréhensibilité du locuteur, c'est-à-dire l'effort que l'interlocuteur doit fournir pour le comprendre, et l'intelligibilité (ce qui est effectivement compris) deviennent donc les paramètres clés de l'évaluation de la prononciation, laissant de côté la perception de l'accent étranger (Derwing, Munro, 2015).

Ce changement de paradigme a été initié dès la fin des années 90, notamment avec la proposition du *Lingua Franca Core* de Jenkins (2000), qui visait à extraire les fondements de la prononciation de l'anglais pour garantir une communication efficace dans un contexte mondialisé, où le statut de natif ne fait plus tellement sens. Parmi les aspects mis en avant par Jenkins, on retrouve les phonèmes à *haut rendement fonctionnel* (Derwing, Munro, 2015), c'est-à-dire les paires de phonèmes qui distinguent une grande quantité de mots, fréquents et de catégories syntaxiques proches. Les consonnes /θ ð ɹ l/ sont ainsi considérées moins importantes que les autres, contrairement à l'aspiration des occlusives en début de mot, ou les initiales et médiales des clusters consonantiques. Les accents lexicaux et phrasiques sont également mis en avant, comme le découpage en groupes de sens, ou encore la longueur vocalique. Envisager l'anglais langue étrangère comme *lingua*

franca a toutefois tardé à se généraliser, et est resté propre à cette langue jusqu'à récemment.

Si la plupart des grilles d'évaluation de la prononciation se résument aux vagues critères « Speaking » et « Fluency » (Piccardo, 2016 : 15), il en existe de plus en plus qui considèrent des aspects de la prononciation plus fin et mieux explicités pour aider l'enseignant dans son évaluation.

Le *volume complémentaire du Cadre Européen Commun de Référence pour les Langues* (CECRL, Conseil de l'Europe, 2018) met en avant l'intelligibilité du locuteur, acceptant un accent qui n'affecte pas la compréhension même au niveau C2, et autorisant l'influence des autres langues connues. Des précisions sont données quant à la maîtrise des traits prosodiques (par exemple au niveau B2 : « *Peut utiliser des traits prosodiques (par ex. l'accent, l'intonation, le rythme) pour faire passer le message qu'il a l'intention de transmettre, mais l'influence des autres langues qu'il/elle parle est notable* » (Conseil de l'Europe, 2018 : 142)), l'accent lexical et le rythme sont également mentionnés dès le niveau A1.

L'échelle de compréhensibilité d'Isaacs et al. (2017) a été spécifiquement conçue pour l'évaluation formative, et se donne pour objectif d'aider les enseignants pour améliorer la compréhensibilité de leurs apprenants en anglais langue étrangère. Elle propose une échelle globale de 6 niveaux et 4 sous-échelles ciblant respectivement la phonétique, la fluence, le vocabulaire et la grammaire ; les deux derniers ayant une importance moindre selon les auteurs. La sous-échelle de phonétique fait ressortir l'importance de la position de l'accent lexical, celle de fluence la localisation des marqueurs d'hésitation (sans indice sur leur type ni ce qu'est une position inappropriée). Les auteurs précisent que leur échelle n'est pas adaptée aux tâches de production de parole trop prédictibles, comme la lecture, les phrases indépendantes ou les dialogues.

Citons une dernière grille, celle que proposent Frost et O'Donnell (2018). Celle-ci aussi est conçue pour évaluer l'anglais, et plus spécifiquement celui des apprenants francophones, toujours dans une optique de et non les placer sur une échelle de distance par rapport à un modèle natif. Elle aussi combine des éléments segmentaux et supra-segmentaux, en donnant toujours plus d'importance aux derniers — d'où son nom de *descripteurs prosodiques*. Les descripteurs sont divisés en 5 catégories : rythme et accent lexical, intonation, réduction syllabique, voyelles pleines, et phonotactique (comprenant entre autres les phénomènes de liaisons, d'élision, ou d'assimilation). Le focus porte principalement sur la qualité et la position de l'accent lexical (comme garant du rythme), la variation de l'intonation et son adéquation aux intentions du locuteur, la qualité des voyelles accentuées et

leur position, et la qualité vocalique (vers une voyelle centrale relâchée en contexte réduit, accompagnée d'une réduction de hauteur, d'intensité et de durée).

On constate donc un réel décalage entre ce qu'évaluent les systèmes automatiques et les objectifs pédagogiques mis en avant aujourd'hui. Si les systèmes automatiques portent le plus souvent sur la reconnaissance des phonèmes en parole lue, les grilles d'évaluation de la prononciation mettent clairement la focale sur les aspects relatifs à la compréhensibilité du locuteur, en particulier à travers le rythme et l'accentuation dans le cas de l'anglais, et donnent moins d'importance aux phonèmes eux-mêmes.

3. Quels critères d'évaluation pour le SELF ?

Inspirés par les descripteurs de compréhensibilité d'Isaacs et al. (2017) et par les descripteurs prosodiques de Frost et O'Donnell (2018), et dans la ligne droite du volume complémentaire du CECRL, nous avons choisi de nous focaliser sur le rythme. Nous nous intéresserons plus précisément à la réalisation de l'accent lexical et de la réduction syllabique, ainsi qu'au découpage du flux de parole par l'utilisation des pauses.

À ce niveau, deux questions se posent : quels paramètres acoustiques mesurer pour caractériser ces phénomènes ? Et comment interpréter les valeurs obtenues (qu'est-ce qui est attendu pour un niveau avancé vis-à-vis d'un niveau débutant) ?

L'accent lexical peut être mesuré sur différentes dimensions : Bonneau et Colotte (2011) utilisent un ratio de hauteur et un ratio de durée entre une syllabe théoriquement accentuée et la syllabe qui suit. Ils proposent également un indicateur du nombre de fois où la hauteur maximale tombe sur la syllabe à accentuer. Notons que leur corpus est constitué seulement de mots isolés avec une structure accentuelle prédéterminée (Oo, oOo et Ooo, où O symbolise une syllabe accentuée et o une syllabe non accentuée). Les mêmes auteurs considèrent qu'une voyelle est réduite si elle dure moins de 40 ms et est au moins deux fois plus courte que la durée moyenne des autres voyelles du mot. Kato et al. (2019) proposent un ratio de durée de référence entre les voyelles de syllabes consécutives, sur la base d'un corpus d'enregistrements de mots provenant de divers dictionnaires en ligne. Truong et al. (2018) utilisent le même corpus, mais se concentrent sur la hauteur et l'intensité pour mesurer la distance des productions d'apprenants par rapport au modèle.

Dans le cas où le texte prononcé n'est pas connu ni segmenté en mots, seules des mesures globales sont utilisées. Par exemple, Kang et Johnson (2018) calculent

le nombre de syllabes accentuées par groupe rythmique (groupe de syllabes entre deux pauses de 200 ms ou plus), le pourcentage de groupes rythmiques contenant au moins une syllabe accentuée, le pourcentage de syllabes accentuées, la variation de hauteur, la hauteur moyenne des syllabes accentuées et non accentuées. L'accentuation est détectée au moyen d'un classifieur automatique qui prend en compte la hauteur et la durée de chaque syllabe, et entraîné sur un corpus de parole lue. Toutefois, à l'inverse des études mentionnées dans le paragraphe précédent, ces études ne permettent pas de savoir si les syllabes accentuées sont les bonnes ou non : elles se limitent à détecter leur présence et mesurer leur quantité dans la parole de l'apprenant.

Nous souhaitons combiner les deux approches : comparer les gabarits accentuels réalisés avec un ou plusieurs gabarits de référence, et permettre cette analyse sur une liste non finie de mots en contexte. La solution envisagée est de coupler une reconnaissance de la parole et un alignement forcé de la transcription au signal pour identifier un certain nombre de mots dans la production de l'apprenant, et comparer le gabarit prosodique de ces mots avec un ou plusieurs gabarits de référence provenant d'un dictionnaire accentuel. Nous ciblerons seulement les mots pleins, et plus précisément les noms communs, les verbes et les adjectifs de deux syllabes ou plus. Le gabarit prosodique de chaque mot sera estimé à partir de la durée des syllabes, de leur hauteur et de leur intensité mais ne sera pas directement comparé à un modèle.

Les pauses silencieuses sont considérées comme telles à partir d'une durée variant entre 100 ms (Kang, Johnson, 2018) et 250 ms (Saito et al., 2022, Fontan et al., 2018). Les pauses pleines, généralement des hésitations résultant en un allongement vocalique, sont souvent annotées manuellement, mais la dernière version du script de De Jong et al. (2021) permet de les détecter automatiquement relativement précisément. Toutefois, toutes les études que nous avons lues se concentrent sur la fréquence des pauses et leur durée moyenne (Saito et al., 2022, Kang et Johnson, 2018, Fontan et al., 2018, Coutinho et al., 2016, Bhat et al., 2010), mais aucune ne semble s'intéresser à leur position dans l'énoncé. C'est pourtant ce point qui est mis en avant dans l'échelle de compréhensibilité d'Isaacs et al. (2017). Nous envisageons donc de coupler la détection des pauses pleines et silencieuses avec une analyse morphosyntaxique de l'énoncé, de manière à observer après et avant quelles catégories grammaticales l'apprenant à tendance à placer ses pauses et hésitations.

Notons toutefois que les grilles d'évaluation donnent peu d'indices quant à ce qui est attendu, et s'en remettent souvent au jugement global et intuitif de l'évaluateur (par exemple : «*Les marqueurs d'hésitation sont parfois placés à des*

jonctions inappropriées (notre traduction[2]*)* », niveau 3 de l'échelle de fluence d'Isaacs et al., 2017). Si l'on souhaite mettre au point un système d'évaluation automatique, il reste donc nécessaire de définir ce qu'est une *jonction inappropriée*, de manière à ce que le système puisse émettre un jugement à partir du positionnement des pauses observées. De même pour l'accent lexical, il n'est pas mentionné de valeur attendue concernant la proportion de syllabes à accentuation non appropriée tolérée aux différents niveaux de compétence de l'apprenant.

Si la plupart des études comparent ce type de valeurs mesurées automatiquement avec un corpus de parole native, ce qui nous intéresse quant à nous, ce n'est pas la distance par rapport à un modèle natif mais plutôt la caractérisation de ce qui participe à détériorer la compréhensibilité de l'apprenant. Nous pouvons alors comparer ces valeurs automatiques avec un jugement humain holistique, qui donne une idée de la compréhensibilité du locuteur. Le corpus d'interactions orales du test certificatif CLES va justement nous permettre de le faire.

4. Méthodologie envisagée pour mettre au point le dispositif d'évaluation

La conception du dispositif diagnostique est envisagée en deux phases :

- Phase exploratoire : observer les patterns de position et de la qualité de réalisation des accents lexicaux, ainsi que du positionnement des pauses chez des apprenants certifiés B1 et B2 ;
- Phase de production : proposer une chaîne de traitement automatique permettant de comparer un nouvel enregistrement de parole avec le modèle B1/B2 développé dans la phase exploratoire.

La première phase est la plus importante : avant d'évaluer, il faut identifier ce qui caractérise la prononciation d'un apprenant à tel ou tel niveau de compétence. Maintenant que les paramètres à mesurer sont choisis, il faut pouvoir calibrer le système sur une échelle de compétence, et voir de quelle manière les valeurs évoluent en fonction des locuteurs et de leur niveau d'anglais. Nous proposons donc une première phase exploratoire pour observer quantitativement quelles sont les tendances d'accentuation et d'utilisation des pauses chez les apprenants.

Pour cela, la certification CLES a accepté de mettre à notre disposition ses enregistrements d'interactions orales du CLES B2. Le CLES permet aux étudiants qui ne valident pas le niveau B2 escompté, de valider le niveau B1, ce qui aboutit à un corpus d'enregistrements d'étudiants certifiés dans l'un ou l'autre niveau, et nous permettra de comparer les valeurs acoustiques observées dans les deux populations. Le corpus CLES dont nous disposons est actuellement constitué de 25 enregistrements audio et 47 locuteurs.

4.1. Description de la tâche de production

Les candidats sont convoqués par binôme pour l'épreuve d'interaction orale. L'examinateur leur donne une fiche contenant le sujet du débat, leur position (pour ou contre) et quelques arguments qu'ils peuvent utiliser. Les candidats ont ensuite 2 minutes de préparation silencieuse pour mettre en forme leur argumentation, puis 10 minutes pour débattre. L'objectif est d'arriver à un compromis. Le corpus actuel est composé de 10 enregistrements de la session 1 du CLES B2 à Grenoble en 2020, et 25 enregistrements de la session 2 du CLES B2 à Grenoble en 2022. Le sujet des premiers était « êtes-vous pour ou contre la e-cigarette ? », celui des derniers était « êtes-vous pour ou contre la généralisation des caméras de surveillance ? ». Pendant la préparation, les candidats disposent d'une feuille de brouillon pour noter leurs idées.

4.2. Critères d'évaluation

Une fois le débat terminé, les candidats sortent de la pièce, et l'examinateur indique si le niveau B2 est acquis dans 8 critères d'évaluation relatifs à la compétence d'interaction orale. Parmi les critères, deux concernent explicitement la prononciation : l'aisance (« Exprime ses idées avec fluidité sans faire de longues pauses (hésitations tolérées) ; Exprime ses idées malgré des pauses pour chercher ses mots ») et la phonologie (« Prononciation et intonation suffisamment claires pour être aisément compris, même si un accent subsiste ; globalement compréhensible malgré l'accent étranger et/ou des erreurs de prononciation »).

Si le niveau B2 n'est pas acquis dans l'un des critères, l'examinateur attribue le niveau B1, ou bien « non validé ». Le candidat doit avoir acquis B2 sur les 8 critères pour valider le niveau à cette épreuve, sinon, le jury final se charge de trancher. C'est également ce jury qui se charge de délivrer le niveau global : B2, B1, ou « non validé ». Sur les 47 locuteurs, 17 sont certifiés B1 et 30 sont certifiés B2.

Pour chaque locuteur, nous disposons donc d'un niveau de compétence global (niveau CLES), d'un niveau pour l'épreuve d'interaction orale, et d'un niveau pour les sous-critères d'aisance et de phonologie.

4.3. Traitements automatiques

Les traitements n'ont fait que commencer, et il ne sera malheureusement pas possible de tirer de conclusions dans cet article. Nous présentons seulement la chaîne de traitement pour chaque enregistrement :

- Détection de la parole et identification automatique du locuteur (outil utilisé : SpeakDiar[3], Bredin et al. 2019) ;
- Reconnaissance automatique de la parole (outil utilisé : AmberScript[4]) ;
- Alignement forcé de la transcription au signal audio (outil utilisé : WebMAUS[5]) ;
- Détection des pauses silencieuses et des pauses pleines (script utilisé : De Jong et al., 2021) ;
- Détection des noyaux syllabiques (idem) ;
- Extraction des paramètres syllabiques (intonation, intensité, durée sur le logiciel Praat) ;
- Analyse morphosyntaxique de la transcription puis récupération du contexte syntaxique gauche et droit de chaque pause (analyseur morphosyntaxique utilisé : SpaCy[6]) ;
- Comparaison du gabarit prosodique des noms communs, verbes et adjectifs avec un dictionnaire de référence.

Jusque-là, seules les 5 premières étapes ont été effectuées. La partie la moins évidente est la reconnaissance de la parole, étant donné qu'il s'agit de dialogues spontanés en L2. Nous avons sélectionné 17 extraits de parole de différents locuteurs et testé 10 systèmes d'ASR[7], de manière à identifier celui qui fonctionne le mieux avec ce type de données. AmberScript est celui qui a obtenu le taux d'erreur mot moyen le plus faible (0,25) et le plus stable entre les différents extraits (écart type de 0,08). C'est malheureusement le plus cher des 8 systèmes propriétaires testés, et son utilisation sera conditionnée par l'obtention des fonds nécessaires ou de la mise en place d'une collaboration entre l'Université Grenoble Alpes et l'entreprise AmberScript.

Après la reconnaissance de la parole, c'est son alignement au signal qui peut parfois donner des résultats erronés, et les analyses qui en découlent s'en retrouvent biaisées. À ce jour, deux outils d'alignement ont été testés : WebMAUS (Schiel, 1999) et EasyAlign (Goldman, 2011), seul le premier a donné des résultats exploitables, mais il reste nécessaire de tester d'autres systèmes (comme P2FA[8] (Yuan, Liberman, 2008), ou encore Gentle[9]) et établir un protocole précis pour quantifier les erreurs d'alignement, et permettre de bloquer la suite des traitements à partir d'un certain seuil d'erreurs pour ne pas biaiser les résultats. La détection des pauses et des noyaux syllabiques semble assez robuste quant à elle et a le mérite de ne pas dépendre de modèle acoustique, elle s'adapte donc tout à fait à la parole non standard des apprenants. Bhat et al. (2010) utilisaient une version antérieure du même script (De Jong, Wempe, 2009) et obtenaient déjà une précision de détection de 90 %.

Une fois la chaîne de traitements terminée, il s'agira de comparer les valeurs obtenues entre les niveaux B1 et B2, et voir si des tendances s'observent, si des phénomènes spécifiques apparaissent en B2 et non en B1, ou vice versa. Il sera intéressant également de regarder si des profils de locuteurs apparaissent, révélant peut-être des stratégies de communication, ou des types de difficultés différents.

Premières conclusions

L'objectif du futur module diagnostique de la prononciation dans le SELF est triple : aider les apprenants à identifier et prendre conscience des phénomènes de prononciation qui impactent la bonne compréhension de leurs productions orales ; mettre à la disposition des enseignants un outil de mesure systématique et automatique de la prononciation fondé sur un principe de compréhensibilité du locuteur ; et évaluer la parole la plus authentique et spontanée possible.

La principale difficulté réside dans l'identification des paramètres acoustiques qui participent à diminuer la compréhensibilité de l'apprenant. Une revue de la littérature sur le sujet nous a permis de cibler certains phénomènes (présence, qualité et position de l'accent lexical ; position des pauses) ; mais si ces phénomènes sont probablement pris en compte dans une évaluation holistique manuelle, il reste nécessaire de faire le lien entre des mesures acoustiques et différents niveaux de compétence (et donc a priori de compréhensibilité).

Il faut donc commencer par mesurer ces valeurs sur des productions de niveaux différents — nous avons choisi les niveaux B1 et B2 — de manière à produire un modèle d'évaluation qui saura interpréter les mesures sur de nouvelles productions. L'idée n'est pas de prédire si l'apprenant a un niveau B1 ou B2, mais plutôt de donner des indices sur les phénomènes mesurés et le niveau dans lequel ces valeurs ont tendance à être observées. À condition que la rétroaction soit facilement interprétable, ce diagnostic pourra servir directement à l'étudiant pour avoir un retour sur sa performance selon des critères de compréhensibilité (et non de « nativité ») ; il pourra également servir aux enseignants pour cartographier les profils de compétences de leurs étudiants, et aider à identifier les exercices à proposer pour y remédier ; enfin ce diagnostic pourrait également être injecté dans un parcours de formation individualisé, qui s'adapterait en fonction des acquis et des difficultés de chaque étudiant. Au-delà de la formation, ce travail permettra également de mieux décrire les phénomènes de prononciation observés chez les apprenants francophones aux niveaux B1 et B2.

Bibliographie

Arias, J. P., Yoma, N. B., Vivanco, H. 2010. « Automatic intonation assessment for computer aided language learning ». *Speech Communication*, vol. 52, no. 3, p. 254-267.

Bada, I., Fohr, D., Illina, I. 2020. *Reconnaissance automatique de la parole : génération des prononciations non natives pour l'enrichissement du lexique*. ATALA ; AFCP, p. 27-35.

Baker, A. A. 2011. ESL teachers and pronunciation pedagogy: Exploring the development of teachers' cognitions and classroom practices. In : J. Levis, K. LeVelle (Eds.). *Proceedings of the 2nd Pronunciation in Second Language Learning and Teaching Conference*, Sept. 2010., Ames, IA: Iowa State University, p. 82-94.

Becker, K., Edalatishams, I. 2019. ELSA Speak – Accent Reduction [Review]. In : J. Levis, C. Nagle, E. Todey (Eds.), *Proceedings of the 10th Pronunciation in Second Language Learning and Teaching Conference*.

Bernstein, J., Cohen, M., Murveit, H., Rtischev, D., Weintraub, M. 1990. Automatic evaluation and training in English pronunciation. ICSLP. 90. 1185-1188.

Bhat, S., Yoon, S.-Y. 2015. « Automatic assessment of syntactic complexity for spontaneous speech scoring ». *Speech Communication*, n° 67, p. 42-57.

Bhat, S., Hasegawa-Johnson, M., Sproat, R. 2010. Automatic Fluency Assessment by Signal-Level Measurement of Spontaneous Speech. Second Language Studies: Acquisition, Learning, Education and Technology.

Bonneau, A., Colotte, V. 2011. *Automatic Feedback for L2 Prosody Learning. Speech Technologies*, Book, 2.

Bredin, H., Yin, R., Coria, J.-M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., Gill, M.-P. 2019. Pyannote.audio: neural building blocks for speaker diarization. ArXiv.

Breitkreutz, J.A., Derwing, T.M., Rossiter, M.J. 2001. « Pronunciation Teaching Practices in Canada ». *TESL Canada Journal*, 19, 51-61.

Burgess, J., Spencer, S. 2000. « Phonology and Pronunciation in Integrated Language Teaching and Teacher Education ». *System*, 28, 191-215.

Chen, L., Zechner, K. 2011. Applying rhythm features to automatically assess non-native speech. in Proc. Interspeech 2011. ISCA, p. 1861-1864.

Conseil de l'Europe, 2001. *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer*.

Conseil de l'Europe, 2018. *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer. Volume complémentaire avec de nouveaux descripteurs*.

Coutinho, E., Hönig, F., Zhang, Y., Hantke, S., Batliner, A., Nöth, E., Schuller, B. 2016. Assessing the Prosody of Non-Native Speakers of English: Measures and Feature Sets. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 1328-1332.

Cucchiarini, C., Strik, H., Boves, L. 2000. « Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms ». *Speech Communication*, n° 30, p. 109-119.

Cucchiarini, C., Strik, H., Boves, L. 2002. « Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech ». *The Journal of the Acoustical Society of America*.

de Jong, N. H., Pacilly, J., Heeren, W. 2021. PRAAT scripts to measure speed fluency and breakdown fluency in speech automatically. *Assessment in Education: Principles, Policy & Practice*, n° 28(4), p. 456-476.

de Jong, N. H., Wempe, T. 2009. « Praat script to detect syllable nuclei and measure speech rate automatically ». *Behavior Research Methods*, n° 41(2), p. 385-390.

Delmonte, R. 2011. Exploring Speech Technologies for Language Learning. In (Ed.), *Speech and Language Technologies. IntechOpen*.

Derwing, T. M., Munro, M. J. 2015. *Pronunciation Fundamentals: Evidence-Based Perspectives for L2 Teaching and Research*. Amsterdam : John Benjamins.

Detey, S., Fontan, L., Pellegrini, T. 2016. « Traitement de la prononciation en langue étrangère : approches didactiques, méthodes automatiques et enjeux pour l'apprentissage ». *Revue TAL,* n° 57(3), p. 15-39.

Didelot, M., Racine, I., Zay, F., Prikhodkine, A. 2019. « Enseignement et évaluation de la prononciation aujourd'hui : l'intelligibilité comme enjeu ». *Recherches en didactique des langues et des cultures, Cahiers de l'Acedle*, n° 16-1. [En ligne] : URL: http://journals.openedition.org/rdlc/4333; DOI: https://doi.org/10.4000/rdlc.4333 [consulté le 15 juillet 2022].

Ding, H., Lin, B., Wang, L., Wang, H., Fang, R. 2020. A comparison of English rhythm produced by native American speakers and mandarin ESL primary school learners. Interspeech 2020.

Duan, R., Kawahara, T., Dantsuji, M., Zhang, J. 2017. Articulatory Modeling for Pronunciation Error Detection without Non-Native Training Data based on DNN Transfer Learning. IEICE Transactions on Information and Systems, vol. E100D, no. 9, p. 2174-2182.

Ellis, N. C., Bogart, P. S. H. 2007. Speech and language technology in education: the perspective from SLA research and practice. In SLaTE-2007, 1-8.

Eskenazi, M. 2009. « An overview of spoken language technology for education ». *Speech Communication*, n° 51(10), p. 832-844.

Evanini, K., Wang, X. 2013. Automated speech scoring for non-native middle school students with multiple task types. Proceedings of Interspeech. 14th Annual Conference of the International Speech Communication Association, p. 2435-2439. Lyon, France.

Field, J. 2005. « Intelligibility and the Listener: The Role of Lexical Stress ». *TESOL Quarterly*, vol.39, n° 3, p. 399-423.

Fontan, L., Le Coz, M., Detey, S. 2018. Automatically measuring L2 speech fluency without the need of ASR: A proof-of-concept study with Japanese learners of french. Interspeech 2018.

Franco, H., Neumeyer, L., Kim, Y., Toledo-Ronen, O. 1997. Automatic pronunciation scoring for language instruction. 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2, 1471-1474 vol.2.

Frost, D., O'Donnell, J. 2018, Evaluating the essentials, the place of prosody in oral production. In :J. Volin (ed.). *Pronunciation of EFL*.

Fu, J., Chiba, Y., Nose, T., Ito, A. 2020. « Automatic assessment of English proficiency for Japanese learners without reference sentences based on deep neural network acoustic models ». *Speech Communication,* n° 116, p. 86-97.

Gilquin, G., Bestgen, Y., Granger, S. 2022. Assessing EFL Speech: A Teacher-Focused Perspective. *Journal of Second Language Teaching and Research*. Volume 9, 2022

Goldman, J.-P. 2011. EasyAlign: An Automatic Phonetic Alignment Tool Under Praat. Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech, 3233–3236.

Goronzy, S., Rapp, S., Kompe, R. 2004. « Generating non-native pronunciation variants for lexicon adaptation ». *Speech Communication,* n° 42(1), p. 109-123.

Isaacs, T. 2016. Assessing Speaking. In : Tsagari, D. & Banerjee, J. (dir.). *Handbook of Second Language Assessment*. Berlin: DeGruyter Mouton, p. 131-146.

Isaacs, T., Trofimovich, P., Foote, J. A. 2017. Developing a user-oriented second language comprehensibility scale for English-medium universities. *Language Testing,* n° 35(2), p. 193-216.

Jenkins, J. 2000. *The Phonology of English as an International Language*. Oxford: Oxford University Press.

Johnson, D. O., Kang, O. 2015. Automatic prominent syllable detection with machine learning classifiers. *International Journal of Speech Technology,* n° 18(4), p. 583-592.

Kang, O., Johnson, D. 2018. « The roles of suprasegmental features in predicting English oral proficiency with an automated system ». *Language assessment quarterly,* 1n° 5(2), p. 150-168.

Kato, T., Truong, Q.-T., Kitamura, K., Yamamoto, S. 2019. Referential Vowel Duration Ratio as a Feature for Automatic Assessment of L2 Word Prosody. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6595–6599.

Kennedy, S., Blanchet, J. Guénette, D. 2017. Teacher-Raters'Assessment of French Lingua Franca Pronunciation. In : Isaacs, T. & Trofimovich, P. (dir.). *Second Language Pronunciation Assessment: Interdisciplinary Perspectives.* Bristol: Multilingual Matters, p. 210-236.

Krashen, S. 2013. « Rosetta Stone: Does not provide compelling input, research reports at best suggestive, conflicting reports on users' attitudes ». *International Journal of Foreign Language Education*, n° 8 (1)

Krashen, S. 2014. « Does Duolingo "trump" university-level language learning? *The International Journal of Foreign Language Teaching* », p. 13-15. [En ligne] : http://sdkrashen.com/content/articles/krashen-does-duolingo-trump.pdf [consulté le 30 juin 2022].

Lee, A., Glass, J. 2015. Mispronunciation Detection Without Non-Native Training Data. in Proc. of Interspeech, p. 643-647.

Levis, J. 2007. « Computer technology in teaching and researching pronunciation ». *Annual Review of Applied Linguistics*, n° 27, p. 184-202.

Li, W., Siniscalchi, S. M., Chen, N. F., Lee, C. 2016. Improving Non-Native Mispronunciation Detection and Enriching Diagnostic Feedback with DNN-based Speech Attribute Modeling. in Proc. of ICASSP, p. 6135–6139.

Loukina, A., Zechner, K., Chen, L., Heilman, M. 2015. Feature selection for automated speech scoring. Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications Association for Computational Linguistics, Denver, CO. p. 12-19.

Moustroufas, N., Digalakis, V. 2007. « Automatic pronunciation evaluation of foreign speakers using unknown text ». *Computer Speech & Language*, n° 21(1), p. 219-230.

Naijo, S., Ito, A., Nose, T. 2021. Improvement of automatic English pronunciation assessment with small number of utterances using sentence speakability. Interspeech 2021.

Neumeyer, L., Franco, H., Digalakis, V., Weintraub, M. 2000. Automatic scoring of pronunciation quality. *Speech Communication*, n° 30(2–3), p. 83–93.

Neumeyer, L., Franco, H., Weintraub, M., Price, P. 1996. Automatic Text-Independent Pronunciation Scoring Of Foreign Language Student Speech. Proc ICSLP. 96.

Nielson, K. B. 2011. « Self-study with language learning software in the workplace: What happens? ». *Language Learning & Technology,* n° 15(3), p. 110-129.

O'Brien, M. 2006. Teaching pronunciation and intonation with computer technology. In L. Ducate & N. Arnold (Eds.), Calling on CALL: From theory and research to new directions in foreign language teaching (CALICO Monograph Series, Vol. 5, p. 127-148.

Pearson Education, Inc. 2022. Versant english test. [En ligne] : https://www.pearson.com/english/versant/our-tests/english-speaking-test.html [consulté le 30 juin 2022].

Pennington, M. 1999. « Computer-aided pronunciation pedagogy: Promise, limitations, directions ». *Computer Assisted Language Learning*, n° 12(5), p. 427-440.

Piccardo, E. 2016. Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Phonological Scale Revision Process Report. [En ligne] : https://rm.coe.int/phonological-scale-revision-process-report-cefr/168073fff9 [consulté le 30 mai 2022].

Saito, K., Macmillan, K., Kachlicka, M., Kunihara, T., Minematsu, N. 2022. Automated assessment of second language comprehensibility: Review, training, validation, and generalization studies. *Studies in Second Language Acquisition*, 1-30.

Schiel, F. 1999. Automatic phonetic transcription of nonprompted speech. Proc. of the ICPhS, 607-610.

Shen, Y., Yasukagawa, A., Saito, D., Minematsu, N., Saito, K. 2021. Optimized prediction of fluency of L2 English based on interpretable network using quantity of phonation and quality of pronunciation. In 2021 IEEE Spoken Language Technology Workshop (SLT) IEEE, p. 698-704.

Tan, T. 2008. *Reconnaissance automatique de la parole non-native*. Thèse de doctorat dirigée par Besacier, Laurent et Castelli, Eric. Grenoble 1.

Truong, Q.-T., Kato, T., Yamamoto, S. 2018. Automatic assessment of L2 English word prosody using weighted distances of F0 and intensity contours. Interspeech 2018.

Walker, R., Low, E., Setter, J. 2021. *English pronunciation for a global world*. Oxford: Oxford University Press.

Wang, X., Zhang, J., Nishida, M., Yamamoto, S. 2015. Phoneme set design for speech recognition of English by Japanese. IEICE Transactions on Information and Systems, E98.D(1), p. 148-156.

Witt S. 2012. Automatic Error Detection in Pronunciation Training : Where we are and where we need to go. Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training, Stockholm, p. 1-8.

Witt, S. 1999. Use of Speech recognition in computer-assisted language learning. Unpublished thesis, Cambrige Uni. Eng. Dept.

Witt, S., Young, S. 2000. « Phone-level pronunciation scoring and assessment for interactive language learning ». *Speech Communication*, n° 30, p. 95-108.

Yoon, S. Y., Bhat, S., Zechner, K. 2012. Vocabulary profile as a measure of vocabulary sophistication. In : *Proceedings of the Seventh Workshop on Building Educational Applications* Montreal, Canada: Association for Computational Linguistics, p. 180-189.

Yuan, J., Liberman, M. 2008. Speaker identification on the SCOTUS corpus. *The Journal of the Acoustical Society of America*, n° 123(5), p. 3878-3878.

Zechner, K., Evanini, K. 2019. *Automated speaking assessment*. London, England: Routledge.

Notes

1. Projet de thèse : *Diagnostic semi-automatique de la parole spontanée en anglais langue étrangère : le rôle du rythme dans la compréhensibilité du locuteur*. Sylvain Coulange, sous la Direction de Monica Masperi, Université Grenoble Alpes.

2. *Hesitation markers are occasionally used at inappropriate junctures*.

3. SpeakDiar : https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/SpeakDiar [consulté le 30 juin 2022].

4. AmberScript : https://www.amberscript.com/ (consulté en juin 2022).

5. WebMAUS : https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/WebMAUSBasic [consulté le 30 juin 2022].

6. SpaCy : https://spacy.io/ [consulté le 30 juin 2022].

7. Dans l'ordre croissant de taux erreur mot : AmberScript, Fraunhofer, Google-us, Google-gb, LSTEnglish-us, LSTEnglish-gb, EML-gb, EML-us, et enfin deux modèles open-source de SpeechBrain (asr-wav2vec2-commonvoice et asr-crdnn-rnnlm-librispeech).

8. P2FA : https://babel.ling.upenn.edu/phonetics/old_website_2015/p2fa/index.html [consulté en juin 2022].

9. Gentle : https://lowerquality.com/gentle/ [consulté en juin 2022].