Traitement de la prononciation en langue étrangère : approches didactiques, méthodes automatiques et enjeux pour l'apprentissage

Sylvain Detey, Lionel Fontan, Thomas Pellegrini

Traitement de la prononciation en langue étrangère : approches didactiques, méthodes automatiques et enjeux pour l'apprentissage

Sylvain Detey* — **Lionel Fontan**** — **Thomas Pellegrini*****

* *SILS, Université Waseda ; Tokyo, Japon – detey@waseda.jp*
** *Archean Technologies ; Montauban, France – lfontan@archean.fr*
*** *Université de Toulouse ; UPS, IRIT ; Toulouse, France – thomas.pellegrini@irit.fr*

RÉSUMÉ. *Cet article, consacré au lien entre traitement automatique et apprentissage de la prononciation en L2, vise à offrir une vue d'ensemble des méthodes et pratiques actuellement en cours dans les deux domaines de référence du sujet (la didactique des langues étrangères et le traitement automatique de la parole), en nous concentrant sur la démarche de correction phonétique, du diagnostic d'erreurs aux procédures de remédiation. L'un des composants les plus novateurs de notre réflexion porte sur les apports de la phonologie de corpus, notamment des corpus oraux d'apprenants de L2. Cette réflexion nous conduit à plaider en faveur d'une approche interdisciplinaire plus riche entre didacticiens et ingénieurs de la parole afin d'encourager le développement des systèmes de correction automatique dans les curricula de L2.*

ABSTRACT. *This article focuses on the link between automatic speech processing and L2 pronunciation learning. It stands as a "position paper" in favor of a more interdisciplinary perspective between the fields of Spoken Language Processing and Second Language Education in the process of designing user-friendly and pedagogically efficient Computer-Assisted Pronunciation Training systems. It offers an overview of current approaches and techniques in both fields (second language pronunciation teaching, from diagnostic to corrective feedback, and automatic pronunciation errors detection and correction). One distinctive aspect of our contribution lies in its connection with the field of corpus phonology, especially oral L2 learners' corpora. In conclusion, we call for more interactions between speech engineers and L2 education specialists to promote the use of such systems in L2 curricula.*

MOTS-CLÉS : *traitement automatique de la parole, prononciation en L2, corpus oraux.*

KEYWORDS: *spoken language processing, L2 pronunciation, speech corpora.*

1. Introduction

Lorsque l'on se penche sur le rôle des avancées techniques en didactique des langues, on doit rappeler le rôle essentiel joué par les phonéticiens au tournant du XX^{e} siècle dans la modernisation de la pédagogie des langues vivantes (le mouvement dit de la « Réforme »). L'apport de la phonétique expérimentale y est essentiel, lui-même rendu possible par les innovations techniques de l'époque, certaines étant transformées en outils pédagogiques (guide-langue, olive nasale, phonographe) (Galazzi, 2002). Au cours du XX^{e} siècle, pourtant, la prononciation a souvent fait figure d'orpheline de la didactique, hormis dans certaines approches méthodologiques, comme la méthodologie structuro-globale audio-visuelle (SGAV, Rivenc 2003). Les développements technologiques portés par l'apprentissage des langues assisté par ordinateur (ALAO) ont réussi à durablement s'installer dans le marché pédagogique, mais leur succès sur le plan phonético-phonologique semble n'avoir pas encore été à la hauteur de ce que les avancées en ingénierie de la parole auraient pu laisser espérer. Une des pistes de recherche les plus récentes dans la réflexion sur l'entraînement à la prononciation assisté par ordinateur (EPAO, en anglais *Computer-Assisted Pronunciation Training*, CAPT) concerne le rôle de plus en plus croissant de la phonologie de corpus, en particulier des corpus d'apprenants, dans le développement des systèmes d'EPAO. Si la linguistique de corpus a en effet été largement mise à profit dans le domaine de l'ALAO depuis les années 1980 en particulier *via* l'utilisation de corpus d'écrits natifs pour le développement de concordanciers et de ressources numériques lexicographiques, ce n'est que plus récemment que les corpus oraux d'apprenants, moins nombreux[1], ont commencé à être exploités en lien avec les technologies de reconnaissance automatique de la parole (RAP) (Carranza *et al.*, 2014). Toutefois, les applications conçues dans cette optique restent encore limitées, tant dans la distinction, parfois complexe, entre ce qui relève de la phonétique et de la phonologie, que du point de vue des besoins et des contingences didactiques. Dans cet article, nous adoptons comme point de départ la didactique des langues étrangères, pour rappeler que les prouesses techniques ne peuvent être suivies de succès pédagogiques qu'à condition d'épouser les besoins contextualisés des apprenants de langue et de leurs enseignants. Nous rappelons dans un premier temps les enjeux d'apprentissage et les grands traits des méthodes classiquement employées dans l'enseignement et la correction de la prononciation (Champagne-Muzar et Bourdages, 1998 ; Lauret, 2007 ; Derwing et Munro, 2015), avant d'examiner l'état actuel des systèmes d'évaluation automatique de la parole et d'EPAO. Sur cette base, nous discutons enfin des défis méthodologiques et techniques qui doivent être abordés par l'ingénierie de la parole si elle souhaite intégrer les avancées les plus récentes dans le domaine de la prononciation en langue étrangère (ci-après L2).

1. Consulter l'inventaire des corpus d'apprenants dans le monde, piloté par S. Granger : https ://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

2. La correction de la prononciation en L2 : histoire, enjeux et méthodes

2.1. *Apprentissage de la parole en langue étrangère et prononciation*

L'histoire de la correction phonétique dans l'enseignement des langues vivantes peut prendre comme point de départ, à la fin du XIX^{e} siècle, la rencontre entre les développements de la phonétique d'une part (notamment *via* la création de l'Association phonétique internationale à l'initiative de P. Passy) et celui de la place de l'oral dans les méthodes d'enseignement des langues d'autre part, avec les précurseurs de ce qui en France deviendra la méthode directe au début du XX^{e} siècle (Galazzi, 2002). L'oralité et la prononciation continuent de garder une place privilégiée dans l'enseignement des langues durant la deuxième moitié du XX^{e} siècle, avec les méthodes audio-orales, puis audiovisuelles, en particulier, dans le cas du français langue étrangère, la méthodologie SGAV, qui profite des avancées à la fois techniques (magnétophone, projecteur d'images fixes, laboratoire de langues) et linguistiques de l'époque (Rivenc, 2003) (utilisation du corpus « *Le français fondamental* », lien avec les travaux sur la rééducation auditive et la linguistique de la parole de P. Guberina). Le SGAV est en effet le cadre méthodologique dans lequel s'insère la méthode verbo-tonale (MVT) de correction phonétique (Guberina *et al.*, 1965), qui a influencé bon nombre d'outils et de pratiques dans le domaine. La priorité étant donnée à l'oral, le travail phonétique constituait un volet majeur de toutes les premières étapes d'apprentissage des méthodes SGAV (desquelles était initialement évacué le support écrit). Avec l'avènement des approches communicatives durant les années 1970, le souci d'exactitude articulatoire des phonéticiens appliqués d'une part, et celui de la fidélité (en perception ou en production) dans la reproduction des patrons énonciatifs des enseignants de la génération suivante d'autre part, sont abandonnés au profit d'une mise en valeur de la compréhensibilité globale, suffisante, selon ses promoteurs, à la réussite d'une communication orale réussie. Tout comme sur le plan lexico-grammatical, la précision phonétique est délaissée au profit de la fluidité (relative) de l'interaction : l'important n'est plus d'avoir une « bonne prononciation » (correspondant implicitement à celle d'un locuteur natif de prestige), mais de se faire suffisamment comprendre. Le principe d' « intelligibilité » (*intelligibility principle*) est donc souvent mis en avant et opposé à celui de « nativité » (*nativeness principle*). Il importe ici de bien distinguer « intelligibilité », « compréhensibilité » et « degré d'accent », puisque, comme le suggère l'étude de Munro et Derwing (1995), même si le degré d'accent est corrélé avec la compréhensibilité perçue et l'intelligibilité, un fort degré d'accent étranger ne diminue pas nécessairement la compréhensibilité ou l'intelligibilité de la parole non native. Nous adopterons donc à ce stade les définitions proposées par Derwing et Munro (2015, p. 175-182) dans leur glossaire[2] : (i) *accent* : aspects de la prononciation qui distinguent les membres de différentes communautés de parole, résultant souvent de différences régionales, ethniques et de classes sociales. Les accents étrangers sont le résultat de l'influence de la langue première (ci-après L1) sur la L2 ; (ii) *compréhensibilité* : le degré d'effort que doit fournir un auditeur pour comprendre

2. Notre traduction.

un énoncé ; (iii) *intelligibilité* : une mesure du degré de compréhension par un auditeur de ce qu'un locuteur a dit. L'intelligibilité est souvent évaluée à travers des transcriptions, des réponses à des questions de type vrai/faux, ou des réponses à des questions de compréhension ; (iv) *fluence* : le degré de fluidité de la parole sans pause ou autres marqueurs de disfluence tels que les faux départs. Il semble que l'enseignement de la prononciation soit aujourd'hui principalement orienté vers l'intelligibilité et la compréhensibilité[3], et à ce titre peu développé dans les formations d'enseignants, hormis chez les spécialistes. Cette carence contraste nettement avec les avancées effectuées dans les différents domaines afférents : apprentissage de la phonologie en L2 (Bohn et Munro, 2007 ; Edwards et Zampini, 2008 ; Colantoni *et al.*, 2015), évaluation de la parole en L2 (Isaacs, 2016 ; Isaacs et Thomson, 2013), effet de l'instruction (Saito, 2012 ; Thomson et Derwing, 2015) ou de la modalité visuelle (Hardison, 2007) sur la prononciation en L2. Dans la section suivante, nous examinons donc brièvement les approches et les techniques employées dans l'enseignement/apprentissage de la prononciation d'une L2, avant de nous interroger sur les apports actuels et à venir du traitement automatique de la parole (TAP) en la matière.

2.2. *La correction phonétique : de l'apprentissage phonologique à la représentation sociale*

Si l'articulation entre traitement automatique de la parole et enseignement/apprentissage de la prononciation en L2 est souvent envisagée sous forme de programme dédié à la correction phonétique (*via* diverses tâches), il semble nécessaire, si l'on souhaite optimiser le système didactique, de comprendre le processus d'acquisition d'un nouveau système phonético-phonologique dans sa globalité, avant de considérer les procédures dans lesquelles le TAP peut intervenir de manière efficace. La première erreur élémentaire consiste à assimiler « prononciation » et « articulation » : l'articulation, c'est-à-dire l'implémentation motrice de séquences phonétiques, n'est que la manifestation ultime d'une composante du traitement, à savoir la production orale d'éléments linguistiques dans une tâche donnée (de la répétition de logatomes conformes aux règles phonotactiques de la langue cible à l'énoncé d'un discours spontané en passant par la lecture à voix haute ou la traduction de phrases écrites par exemple). Or, la maîtrise de la prononciation consiste avant tout à acquérir un système phonético-phonologique qui concerne tout aussi bien le versant perceptif que productif de la compétence linguistique, et qui, tant au niveau segmental que suprasegmental, constitue le cœur de l'apprentissage morpholexical (paires minimales, alternances morphophonologiques) et discursif (en particulier au niveau pro-

3. Il faut également noter que, dans le domaine du traitement de la parole, la distinction entre intelligibilité et compréhensibilité est souvent associée à celle entre transfert de forme (phonético-acoustique) et transfert de sens. L'intelligibilité est généralement évaluée par des tâches de transcription ou de répétition de mots, tandis que la compréhensibilité est testée dans des tâches plus proches de situations de communication réelles et impliquant des processus cognitifs de plus haut niveau (cf. Fontan *et al.*, 2015 ; Wilson et Spaulding, 2010)

sodique). C'est en effet sur l'apprentissage phonético-phonologique que repose l'apprentissage du lexique, et partant de la L2 : la fonction distinctive vise avant tout, et *in fine*, à produire du sens, auquel seront associées certaines formes phonético-phonologiques. Ajoutons à cela la dimension prosodique audiovisuelle attitudinale (Rilliard *et al.*, 1988 ; Lu, 2015) ainsi que la question de l'impact social des accents (régionaux ou étrangers) et des représentations sociophonétiques qui leur sont associées (Moyer, 2013) tant chez les natifs que les non-natifs et l'on réalise à quel point les enjeux de la « prononciation » sont bien plus larges et pluriels qu'il n'y paraît. On doit donc distinguer au moins quatre angles d'approche (Detey, 2016) : (1) enseignement de la prononciation (présentation du, exposition au, et activités d'apprentissage du système phonético-phonologique de la L2, de manière plus ou moins intégrée aux autres composantes du matériel langagier à apprendre) ; (2) correction phonologique (activités de remédiation au niveau phonologique relatives à des productions inappropriées (ex. confusions phonémiques) pouvant affecter l'intelligibilité, voire la compréhensibilité) ; (3) correction phonétique (activités de remédiation au niveau phonétique relatives à des productions intelligibles mais pouvant affecter la compréhensibilité et porteuses d'un accent potentiellement non désiré par l'apprenant) ; (4) développement phonopragmatique (élargissement des habiletés sociophonétiques et phonostylistiques de l'apprenant lui permettant de reconnaître et de produire différents types de parole, par exemple socialement indexés). Si ces quatre angles peuvent être agencés de différentes manières, c'est sur la question de la « correction » que repose pour l'essentiel la connexion avec le traitement automatique de la parole. Cette notion renvoie en effet à un standard, à partir duquel sont évaluées des déviations. Cette évaluation peut s'exprimer de différentes manières (correct/incorrect, acceptable/non acceptable, natif/non natif, conforme à la cible/non conforme à la cible, etc.) et ce de manière binaire (bon/mauvais) ou scalaire (sur une échelle graduée). Or, contrairement à la grammaticalité, sur les plans morphologique et syntaxique notamment, les jugements de bonne formation sur le plan phonético-phonologique sont moins évidents, dans la mesure où la norme en la matière est beaucoup plus difficile à circonscrire (Laks, 2002), et où coexistent souvent plusieurs normes. La notion de « français standard », par exemple, a pour cette raison cédé la place à celle de « français de référence » parmi les sociolinguistes (Morin, 2000 ; Lyche, 2010). Néanmoins, d'un point de vue pédagogique et dans les représentations non expertes, la notion de standard fait toujours sens (Detey *et al.*, 2016a), et les systèmes de correction automatique ont besoin de tels repères. La notion d' « erreur », doit ainsi être interprétée de manière nuancée, en particulier quand on sait que, outre la variation inter- et intra-individuelle existant dans les communautés natives (sur le plan diatopique, mais également diastratique et diaphasique), le mécanisme d'apprentissage d'une L2 consiste précisément en l'évolution, progressive et instable, de microsystèmes (l'interlangue/les interlangues) qui vont graduellement se rapprocher d'une version stabilisée de la L2 (Vogel, 1995). Sur le plan phonético-phonologique, la variété de l'apprenant à un stade donné (l'« interphonologie ») va donc présenter certaines caractéristiques, qui pourront être qualifiées de déviances, par rapport au système cible. On peut alors proposer une typologie de ces déviances, en commençant par celle de Moulton (1962) : *phonémiques* (inventaires phonémiques différents entre la langue source (ci-après L1) et la L2, ex. absence de /y/ en arabe stan-

dard moderne) ; *phonétiques* (équivalences phonémiques mais non phonétiques entre L1 et L2, ex. réalisations prototypiques différentes de /u/ en français (postérieure et arrondie) et en japonais (plus centralisée et non arrondie)) ; *allophoniques* (équivalences phonémiques et phonétiques, mais non allophoniques, c'est-à-dire pas pour tous les allophones, entre L1 et L2, ex. /s/ réalisé [s] en français et en japonais, mais palatalisé [ç] devant /i/ en japonais (et typiquement catégorisé comme /ʃ/ par des auditeurs francophones natifs)) ; *distributionnelles* (équivalences d'unités mais pas de distribution entre L1 et L2, ex. absence de groupe /sp/ à l'initiale de mot en espagnol ou dévoisement des obstruantes voisées en coda syllabique en allemand). Sur le plan segmental, on distingue généralement les insertions, les effacements, les substitutions et les distorsions (Derwing et Munro, 2015, p. 58), et on peut également adopter une approche infrasegmentale des déviances (*via* des matrices de traits distinctifs, par exemple de voisement, souvent en lien avec la dimension phonotactique et, sur le plan phonétique, les phénomènes de coarticulation). Il faut bien sûr ajouter à cela les déviances d'ordre suprasegmental, en particulier accentuel, rythmique et intonatif, puisque le domaine syllabique est lié à la dimension phonotactique (distributionnelle). Dans une perspective pédagogique, la hiérarchisation de ces déviances est essentielle et peut être effectuée selon plusieurs critères : leur « gravité » (interprétée par exemple à travers la notion de charge fonctionnelle (*Functional Load*[4]) qui renvoie notamment à la productivité d'un contraste, voir par exemple Brown (1988), Munro et Derwing (2006), Kang et Moran (2014)), leur fréquence d'occurrence, ou encore leur impact sociolinguistique (lequel doit tenir compte de la communauté native impliquée, ex. Paris *vs* Montréal dans le cas du français). Telle déviance sera jugée plus sérieuse que d'autres et donc à traiter en priorité, certains éléments étant plus à même de conférer un « accent » que d'autres (Vaissière et Boula de Mareüil, 2004 ; Vieru *et al.*, 2011). Le rôle du système correctif (enseignant humain ou système automatique) est donc d'aider l'apprenant à réduire ces déviances. Pour ce faire, quatre étapes sont à considérer :

– pronostic : établi *a priori* sur la base du profil sociophonétique de l'apprenant (L1, autres langues connues, séjours en milieu homoglotte, etc.) ;

– diagnostic : établi sur la base des productions effectives de l'apprenant. Il s'agit donc d'une démarche d'évaluation (plutôt formative que sommative), dont la problématisation est au cœur du présent article ;

– remédiation : proposée sur la base d'une expertise dans les domaines du traitement de la parole en L2 et de la didactique de la prononciation ;

4. Derwing et Munro la définissent ainsi (2015, p. 178) : « *A measure of the « work » done by a speech sound in keeping minimal pairs apart* ». Derwing et Munro (2015, p. 75) suggèrent en particulier de prendre en compte les facteurs suivants dans le calcul de la charge fonctionnelle : (i) la fréquence de type (*type frequency*) : le nombre de paires minimales distinguées par les deux segments (plus il y a de paires, plus la charge est élevée) ; (ii) la fréquence d'item (*token frequency*) : la fréquence d'occurrence des mots de la paire (si les deux sont fréquents, la charge est la plus élevée, si l'un des deux est rare et l'autre fréquent, la charge est moins élevée, si les deux sont rares, la charge est la plus faible) ; (iii) la catégorie syntaxique des mots de la paire : si les deux mots appartiennent à la même catégorie (*part of speech*), la charge est plus élevée.

– renforcement (ou systématisation, automatisation, routinisation) : proposé sur la base d'une expertise dans le domaine de l'apprentissage d'une L2 et visant à stabiliser et automatiser le couplage acoustico-articulatoire établi lors de la remédiation.

Avant de nous pencher sur la question centrale de l'évaluation (et donc du diagnostic), examinons brièvement les démarches et les méthodes disponibles employées en didactique de la prononciation pour l'étape de remédiation, qui s'articulent toutes autour de l'axe production/perception : (a) *la méthode articulatoire*, qui, à l'aide de représentations visuelles, d'instructions explicites et d'activités sensori-motrices, insiste sur le positionnement des articulateurs correspondant aux descriptions phonétiques du modèle à suivre ; (b) *la méthode des oppositions phonologiques*, qui, à l'aide d'exercices centrés sur les paires minimales, cherche à faire acquérir, tant en production qu'en perception, le système phonologique de la L2 ; (c) *la méthode perceptive*, qui, à l'aide de diverses techniques centrées sur l'*input* fourni aux apprenants, insiste davantage sur la réalisation de cibles acoustiques acceptables que sur l'articulation *per se*, reposant sur le principe de primauté de la perception sur la production (une production erronée étant, généralement, la conséquence d'une perception erronée). Ce dernier principe, étayé par de nombreuses études montrant l'impact réussi de l'entraînement perceptif sur la production (Akahane-Yamada *et al.*, 1996 ; Huensch, 2016), y compris à l'aide d'EAPO (Thomson, 2011), a été particulièrement développé en Europe dans le cadre de la MVT : sur la base de la production déviante de l'apprenant (ex. en tâche de répétition), l'enseignant établit un diagnostic à l'aide de principes de classification psycho-acoustiques pédagogiquement orientés (Calbris, 1969 ; Intravaia, 2000 ; Renard, 2002), puis ajuste l'*input* à travers différent procédés, dont font partie les composantes kinésique et posturo-gestuelle, de manière à conduire l'apprenant à réaliser la cible acoustique perceptivement visée[5]. La MVT, s'appuyant sur la métaphore du crible phonologique (Troubetzkoy, 1949), était liée en audiométrie à l'orthophonie et à la rééducation auditive des malentendants, et avait déjà été investie technologiquement avec les appareils « Suvag », constitués de filtres et d'amplificateurs, adaptés à l'usage des enseignants de langue en « Suvag Lingua » (Guberina, 1973)[6]. Ainsi, lorsque l'on examine l'évolution des pratiques dans l'enseignement de la prononciation d'une L2, on observe le passage d'une conception minimalement phonémique de la prononciation à une vision plus large, orientée vers le discours et incluant les traits segmentaux et prosodiques ainsi que la qualité de la voix (*voice setting*). L'enseignement de la prononciation devrait être inclus dans des tâches visant la création de sens référentiel et interactionnel, et non dans de simples activités formelles de vocalisation de mots ou de phrases (Pennington et Richards, 1986). L'intégration équilibrée de la forme et du sens lors du travail sur la prononciation dans des activités communicatives, ainsi que l'impact de l'instruction durant ces activités, ont été clairement mis en avant par plusieurs chercheurs (Isaacs, 2009 ; Saito, 2012), réhabilitant par exemple l'activité de répétition à condition qu'elle soit signifiante (*meaningful repetition* plutôt

5. Voir le site de formation à la MVT réalisé par Michel Billières et son équipe : http ://w3.uohprod.univ-tlse2.fr/UOH-PHONETIQUE-FLE/index.html

6. Voir également : http ://www.suvag.com/

que *rote learning*, Trofimovich et Gatbonton, 2006). Dans le domaine de l'EPAO, après un certain regain d'intérêt pour le rôle de la technologie à la fin des années 1990 (Ehsani et Knodt, 1998 ; Chun, 1998, sur la prosodie), on peut s'interroger aujourd'hui sur l'état des dispositifs, quinze ans après qu'ont été posées certaines grandes orientations pour les systèmes informatisés d'aide à la prononciation (Pennington et Richards, 1999 ; Eskenazi, 1999, plus spécifiquement dans le cas de la reconnaissance automatique de la parole pour l'entraînement à la prononciation). Comme le suggèrent Derwing et Munro (2015, p. 121-122), tout apprenant de langue étrangère souhaitant améliorer sa prononciation pourrait rêver d'un système qui pourrait[7] : « (a) évaluer l'intelligibilité et la compréhensibilité de la parole automatiquement ; (b) identifier les problèmes prosodiques et segmentaux dans la production qui compromettent l'efficacité de la communication ; (c) empêcher de commettre des erreurs humiliantes, et épargner l'embarras d'une correction en personne ; (d) fournir une série d'exercices intéressants avec du *feedback* aidant à résoudre les difficultés identifiées en (b) et (c) ; (e) contrôler l'apprentissage, en fournissant des mesures régulières des progrès effectués et une évaluation des habiletés de prononciation à la fin de l'instruction ». Certaines des questions essentielles qui se posent aujourd'hui semblent donc être les suivantes :

– Les méthodes actuelles d'évaluation automatique de la parole permettent-elles d'intégrer les notions d'intelligibilité et de compréhensibilité, et surtout de moduler la notion de « correction » en intégrant des seuils d'acceptabilité suffisamment précis du point de vue de l'articulation entre les niveaux phonétique et phonologique, tout en incluant une dimension sociolinguistique ?

– Les systèmes actuels d'évaluation automatique de la parole permettent-ils d'optimiser les procédures d'évaluation en tenant compte des profils (essentiellement linguistiques) des apprenants-utilisateurs ?

– Les outils actuels d'évaluation automatique de la parole à des fins pédagogiques offrent-ils un contenu dont le format et les fonctionnalités correspondent aux connaissances actuelles dans le domaine de l'enseignement/apprentissage de la prononciation en L2 ?

Préalablement à une discussion plus fine de ces questions et de leurs enjeux, nous offrons dans ce qui suit une vue d'ensemble des procédés actuels dans le domaine de la reconnaissance automatique de la parole à finalité corrective.

3. Le traitement automatique de la parole appliqué à l'enseignement et à l'apprentissage de la prononciation en langue étrangère

3.1. *Introduction à la reconnaissance automatique de la parole*

La reconnaissance automatique de la parole (RAP) a pour fonction de convertir un signal audio de parole en une séquence de mots correspondant au message linguistique sous-jacent. Cet objectif est le plus souvent décrit en termes probabilistes

7. Notre traduction.

par les ingénieurs de TAP. Ainsi, selon la formulation « classique » introduite il y a quarante ans, au milieu des années 70, il s'agit de trouver la séquence de mots M la plus probable étant donné un signal de parole S. Plus précisément, la RAP cherche à identifier la suite de mots la plus probable qui maximise le produit $P(S|M) * P(M)$ où $P(S|M)$ est la probabilité conditionnelle, appelée *vraisemblance*, d'observer le signal S à partir de la séquence de mots M, et $P(M)$ est la probabilité *a priori* d'observer la séquence de mots M. Cette dernière est généralement estimée à l'aide d'un modèle de langage statistique construit à partir de grands corpus de textes. Autrement dit, la reconnaissance d'une séquence de mots dépend à la fois de sa vraisemblance acoustique calculée à partir du signal de parole, et de la probabilité d'apparition de la suite de mots considérée. Dans un contexte où la probabilité d'occurrence d'un mot est très élevée, par exemple pour le mot « République » après « président de la », la reconnaissance du mot dépendra très peu de sa réalisation acoustique. La RAP repose donc sur des modèles statistiques acoustiques et syntaxiques. Afin de modéliser les aspects phonético-phonologiques de la langue, les systèmes de RAP utilisent tout d'abord des lexiques contenant des variantes de prononciation. À un mot sera associé de 1 à *n* variantes de prononciation, représentées par des suites de phones. Un lexique peut par exemple inclure, pour un mot, des variantes diatopiques (par exemple à l'adjectif « petite » correspondront les formes [pətit], [ptit] et [pətitə]) ou des variantes résultant de phénomènes de sandhi, comme la liaison (par exemple pour le pronom personnel « elles », le lexique contiendra la forme [ɛlz] pour reconnaître des énoncés comme [ɛlzaʁiv] - « elles arrivent »[8]). Les phones (réalisations de phonèmes) sont, quant à eux, représentés par des modèles acoustiques. Ces modèles acoustiques ont longtemps été fondés sur des mélanges de Gaussiennes (*Gaussian Mixture Models*, GMM) qui tentent de capturer la variabilité de paramètres acoustiques extraits sur des fenêtres de 20 ms de signal. Ces GMM sont associés à des chaînes de Markov cachées pour modéliser les transitions entre les phones (Rabiner, 1989). Progressivement, des modèles hybrides sont apparus, en remplaçant les GMM par des réseaux de neurones artificiels pour profiter des bonnes capacités de classification discriminante de ces derniers (Meinedo *et al.*, 2003). Le nouvel état de l'art en RAP repose maintenant sur des réseaux de neurones dits profonds (DNN pour *Deep Neural Networks* et CNN pour *Convolutional Neural Networks*) qui ont la capacité d'apprendre leurs propres représentations des données de manière hiérarchique à travers leurs nombreuses couches denses ou de convolution successives, particulièrement efficaces pour la tâche de classification qu'est la modélisation acoustique de phones (Abdel-Hamid *et al.*, 2012). Les systèmes les plus récents proposent même des réseaux de neurones *end-to-end* qui réalisent la modélisation acoustique et linguistique entière en se passant d'un modèle de langage externe (Amodei *et al.*, 2015). Dans le cadre d'une application d'EPAO qui évalue la prononciation au niveau phonétique, la vraisemblance calculée par un système de RAP est très souvent utilisée comme un score à comparer à un seuil pour

8. Dans un lexique de prononciations de RAP, le phone [z] sert à définir une variante de prononciation du premier mot, « elles », dans un souci d'économie. S'il était rattaché au second mot, « arrivent », il faudrait le rattacher à toutes les prononciations des mots qui sont susceptibles de suivre « elles » et commençant par une voyelle.

décider de la qualité d'une réalisation phonétique. De plus, en EPAO, la situation la plus commune est celle où l'apprenant répète ou lit une phrase donnée. Dans ce cas, la séquence de mots et de phones attendue est connue à l'avance et un système de RAP est utilisé pour réaliser un alignement phonétique entre la séquence de phones attendue et le signal acoustique. Les scores de vraisemblance obtenus lors de cet alignement sont utilisés par de nombreux algorithmes de détection d'erreurs pour qualifier la prononciation de locuteurs L2.

3.2. *La détection automatique d'erreurs de prononciation*

À ce jour la majorité des travaux en détection automatique ont porté sur l'identification d'erreurs segmentales – c'est-à-dire de phonèmes « mal réalisés » par des apprenants de L2 (Eskenazi, 2009 ; Montacié et Caraty, 2015 ; Witt, 2012). À ce champ de recherche dénommé *individual error detection* en anglais s'ajoutent les travaux sur l'identification et la caractérisation d'erreurs à un niveau plus large – généralement de l'ordre du mot ou de la phrase – champ de recherche que l'on désigne sous le terme plus générique de *pronunciation assessment* (Eskenazi, 2009). Ces outils sont pour la plupart développés à partir de corpus d'enregistrements de locuteurs L2 annotés par des experts. C'est ce que l'on appelle l'apprentissage supervisé : le système « apprend » à identifier des erreurs de prononciation en généralisant à partir d'exemples qui lui ont été fournis. Les annotations d'experts (des informaticiens spécialisés dans le traitement de la parole, des phonéticiens ou des enseignants de langue étrangère) servent alors de référence (*groundtruth evidence*) pour l'entraînement du système.

3.2.1. *La détection d'erreurs segmentales*

Pour la détection d'erreurs segmentales la mesure la plus répandue est celle du *Goodness of Pronunciation* (GOP – Witt, 1999). L'idée générale du GOP est d'évaluer à quel point un phone produit se rapproche d'un modèle acoustique natif, en utilisant les statistiques de confiance issues de l'alignement automatique des paroles prononcées par le locuteur L2 (pour les différentes mesures de confiance utilisées voir Hu *et al.*, 2013). À partir d'une transcription phonétique attendue de l'énoncé, deux phases de reconnaissance automatique sont réalisées :

– une phase d'alignement forcé. Le modèle de langage du système de RAP correspond à l'énoncé cible prononcé par l'apprenant (mot ou phrase). Le système cherche donc à délimiter temporellement les phones cibles sur le signal de parole ;

– une phase d'alignement libre. Le système est libre de reconnaître n'importe quelle suite de phones.

Les alignements forcé et libre sont enfin comparés afin de produire un score par phone cible. Si un phone cible est bien reconnu lors de l'alignement libre, ce score sera nul. Sinon, les mesures de confiance lors de l'alignement du phone cible dans les deux phases (forcée et libre) sont traduites en une distance. Plus le score de GOP est élevé, plus cette distance est grande. Des seuils peuvent enfin être définis pour

caractériser une prononciation comme correcte ou erronée. De nombreux auteurs ont élaboré des variantes du GOP et d'autres mesures issues des scores de confiance de RAP. Par exemple le *Forced GOP* (F-GOP) limite la reconnaissance libre de phones aux intervalles temporels définis par l'alignement forcé ; cet algorithme a démontré de meilleurs résultats que la version originale du GOP (Luo *et al.*, 2009). Il est également reconnu que l'élaboration de systèmes intégrant des informations sur la L1 permet d'obtenir de meilleures performances, même si cela est moins adapté à une commercialisation de masse (Witt, 2012). Ces informations sont souvent relatives aux erreurs des apprenants et à leur contexte d'occurrence (Wang et Lee, 2012b ; Wang et Lee, 2012a ; Laborde *et al.*, 2016 ; Li *et al.*, 2016). D'un point de vue purement technique, l'utilisation de DNN pour la reconnaissance automatique permet également une détection plus précise des erreurs de prononciation (Hu *et al.*, 2013 ; Hu *et al.*, 2015). Enfin, il a été démontré que l'utilisation d'algorithmes classifieurs comme les machines à vecteurs de support (*Support Vector Machines* – SVM) ou de réseaux neuronaux à partir de données acoustiques brutes, de scores de confiance de RAP ou bien de cartes d'activation pouvait donner de meilleurs résultats que le GOP (Wei *et al.*, 2009 ; Hu *et al.*, 2015 ; Pellegrini *et al.*, 2016).

3.2.2. *L'évaluation globale de la prononciation (*overall pronunciation assessment*)*

Il existe plusieurs façons d'évaluer la prononciation à un niveau supraphonémique. La première solution consiste à moyenner les scores obtenus localement pour chaque phone (Eskenazi, 2009 ; Hu *et al.*, 2013 ; Chen et Jang, 2012). La seconde méthode est de considérer un faisceau de paramètres phonémiques et prosodiques. Les n différents paramètres considérés vont définir un espace à n dimensions dans lequel il est alors possible de représenter la variabilité des productions orales natives, et donc de pouvoir calculer la distance qui les sépare des productions d'apprenants de L2 (Witt, 2012). Généralement cette distance est exprimée sur une échelle de 1 à 5, allant d'une prononciation totalement inintelligible à une prononciation pouvant être jugée comme native. Les scores de prononciation globale peuvent être calculés pour des mots ou des phrases (Chen et Jang, 2012) ou bien pour l'ensemble d'énoncés produits par un locuteur dans le cas d'une application pour l'évaluation automatique de la compétence orale (Yuan et Liberman, 2016). Cette section n'a pas pour prétention d'effectuer une présentation exhaustive des différents indices phonémiques et prosodiques utilisés pour l'évaluation globale de la prononciation. Si toutefois nous devions citer les indices les plus souvent utilisés, au niveau phonémique nous retrouverions le GOP (ou directement les taux de confiance de systèmes d'alignement automatique), des informations spectrales (formants vocaliques) ou de durée (durées vocaliques, patterns n-grammes de durées phonémiques), et des résultats de classificateurs fonctionnant sur des paires phonémiques ou des catégories phonologiques (Witt, 2012). Du côté prosodique, les indices fréquemment utilisés concernent : *le débit* (ex. nombre de phonème(s) / pseudo-syllabe(s) / mot(s) par seconde ou par minute) ; *la fluence* (ex. fréquence et longueur des pauses) ; *l'intonation* (pente, maximum, écart-type du f_0) ; *la qualité de la réalisation des accents* (ex. quantificateurs de la proximité acoustique entre syllabes accentuées et inaccentuées) ; *l'intensité de réalisation des mots*

(minimum, maximum et moyenne). Cette liste n'est nullement exhaustive, et l'utilisation d'autres paramètres pour l'évaluation automatique de la prononciation est à l'étude – citons par exemple la variabilité de la durée de réalisation des unités phonétiques (Black *et al.*, 2015) ou encore la qualité de la réalisation des frontières interphones (Yuan et Liberman, 2016). Enfin, ces mesures sont parfois combinées à des indicateurs de correction d'autres niveaux linguistiques tels que le lexique et la syntaxe pour prédire la compétence plus globale d'expression orale des apprenants (Qian *et al.*, 2016 ; Tao *et al.*, 2016).

3.3. *La génération de* feedback

La détection d'erreurs n'est qu'une première étape dans la création d'un outil d'entraînement à la prononciation. Pour être efficace, un système complet d'EPAO doit se servir des éléments de diagnostic apportés par le module de détection d'erreurs pour à terme proposer des éléments de correction appropriés. Dans cette optique et selon Hansen (2006), le système doit répondre à quatre exigences : établir des éléments de caractérisation qualitative (1) et quantitative (2) sur les éléments mal prononcés (le diagnostic), et fournir en retour à l'apprenant des informations sous une forme compréhensible (3) et lui permettant de corriger sa production (4) : le *feedback*. La génération automatique de *feedback* a longtemps été guidée par ce qu'il était possible de faire techniquement bien plus que par des considérations d'ordre didactique (Neri *et al.*, 2002). L'implémentation peut-être la plus représentative d'un *feedback* difficilement compréhensible par l'apprenant est la visualisation de la forme d'onde correspondant au signal de parole prononcé, représentée à côté de la forme d'onde du signal cible. Si l'apprenant peut avoir une idée de la distance entre les deux productions, avec ce genre de représentation il n'a aucune information qualitative sur ce qui caractérise sa prononciation par rapport au modèle natif (Ai, 2013). Il risque donc de répéter et de fossiliser ses erreurs plutôt que de les corriger (Eskenazi, 2009). Ce type de *feedback* est présent dans des didacticiels actuels, comme par exemple dans les versions les plus récentes du logiciel TELL ME MORE (2013). Pour aller plus loin, des études ont porté sur la production d'indices utiles à l'autocorrection de la prononciation. En plus des informations purement quantitatives – c'est-à-dire des simples scores de « bonne prononciation » – les systèmes ainsi élaborés profitent par exemple des techniques d'alignement pour fournir à l'apprenant des informations sur l'endroit exact de son erreur, que ce soit à l'échelle du phone (ex. dans le système Fonix Talk SDK, 2016) ou bien du mot (Saz et Eskenazi, 2012). De même, en partant de l'idée que pour être efficace un système d'EPAO doit aider l'apprenant à mieux percevoir les différences entre sa prononciation et la prononciation native (Witt, 2012), d'autres types de *feedback* ont été élaborés afin de fournir à l'apprenant des éléments informatifs ou prescriptifs. Le premier type de *feedback*, et probablement le plus utilisé, est le *feedback* de type articulatoire. Globalement, il s'agit soit de montrer à l'apprenant un modèle du mouvement cible à réaliser pour un son de parole (Miyakoda, 2013), soit de lui montrer son propre mouvement à partir d'algorithmes d'inversion articulatoire (Hueber, 2013). Le feedback *auditif* permet aussi de faire prendre conscience à l'ap-

prenant des différences entre sa production et un modèle natif (Ai, 2013) par divers biais :

– l'emphase : la synthèse de la parole permet par exemple de faire un focus sur l'endroit où l'apprenant a fait une erreur (Meng *et al.*, 2012) ;

– l'exagération : Lu *et al.* (2002) exagèrent par exemple les trois paramètres acoustiques de l'accentuation pour mieux faire percevoir le contraste entre syllabes accentuées et inaccentuées aux apprenants ;

– la transposition prosodique : des paramètres prosodiques (ex. les contours intonatifs) de l'énoncé cible sont transposés sur l'énoncé de l'apprenant. Ce dernier peut donc entendre l'énoncé cible avec sa propre voix. Ce type de technique repose sur des études ayant montré que le *feedback* auditif était d'autant plus efficace que la voix utilisée était proche de celle de l'apprenant (Eskenazi, 2009).

Il faut néanmoins souligner que la plupart des systèmes automatiques se sont focalisés sur la détection d'erreurs et non sur le *feedback* car ce n'est plus problème purement d'ingénierie, ce qui est donc plus difficile à appréhender pour les chercheurs en TAP. À noter également que la majorité des systèmes automatiques abordent principalement les difficultés d'ordre segmental, et bien moins celles d'ordre suprasegmental. Ainsi, comparativement, assez peu d'entre eux se sont investis dans la correction de l'intonation qui est pourtant l'un des aspects prioritaires à traiter dans l'enseignement/apprentissage d'une L2. Les travaux en la matière sont davantage le fait de linguistes, par exemple le logiciel WinPitch (Martin, 2004, http ://www.winpitch.com/), décliné en plusieurs versions : Pro W8 pour la recherche en prosodie, LTL W8 pour les enseignants avec possibilité de créer des leçons et des tests de prononciation, LTL simple, pour les apprenants, avec système d'alignement automatique pour la détection des erreurs et des fonctions de morphing prosodique. Ce système permet de visualiser les courbes intonatives produites par l'apprenant, avec surlignage coloré, d'ajuster la vitesse de lecture, de comparer par alignement automatique le modèle de l'enseignant et la production de l'apprenant, etc.

4. Perspectives linguistiques et didactiques sur le TAP appliqué à l'acquisition d'un nouveau système phonético-phonologique

Au vu des différentes études mentionnées dans la section précédente, on peut apprécier la profusion d'innovations en EPAO, rendues possibles entre autres par les avancées dans le domaine de la RAP. Le chantier d'intégration reste néanmoins ouvert, dans la mesure où certains travaux, notamment en linguistique de corpus, sont eux-mêmes assez récents (Durand *et al.*, 2014), mais aussi parce que, pour une performance optimale, la grande majorité des systèmes d'EPAO restent dépendants des systèmes L2/L1 qu'ils ciblent (Witt, 2012). Trois pistes en particulier doivent être discutées : (i) celle des normes natives prises comme références pour l'évaluation des productions d'apprenants, (ii) celle des apports des corpus d'apprenants, (iii) celle du

format et du *feedback* offerts par les produits pédagogiques dérivés des systèmes de reconnaissance automatique.

4.1. *De la prise en compte des normes et de la variation : les corpus natifs et la perception des natifs*

Puisque l'évaluation des productions d'apprenants s'effectue toujours vis-à-vis d'un modèle de référence, on doit s'interroger sur la portée du modèle adopté, dans la mesure où la (socio)linguistique de la parole contemporaine a parfaitement intégré la difficulté de définir précisément ce que pourrait ou devrait être « la » norme de l'oral sur le plan de la prononciation (Detey *et al.*, 2016a). Il importe sans doute de rappeler ici la distinction entre normes objective, subjective et prescriptive, mais aussi d'insister sur celle entre les niveaux phonétique et phonologique, dans la mesure où, si le dernier présente un plus grand degré de stabilité et d'accord entre locuteurs, les correspondances entre catégories phonologiques et espaces phonétiques sont en revanche sujettes à un certain degré de variation entre locuteurs/auditeurs et de variation interdialectale, en particulier lorsqu'il y a un décalage entre trait phonétique et pertinence phonologique. D'un point de vue acoustique, pour le système vocalique du français de référence par exemple, même s'il est courant dans le domaine phonétique de se référer aux données du groupe Calliope (1989), leurs limites sont connues (liées notamment aux contextes limités d'élicitation) et d'autres chercheurs ont poursuivi la tentative de fournir des valeurs de référence à partir d'autres données (Gendrot et Adda-Decker, 2005), notamment à des fins didactiques (Georgeton *et al.*, 2012). Néanmoins, par-delà l'éclairage majeur sur la réalité et l'importance de la variation phonético-phonologique inter- et intralocuteurs apporté par la phonologie de corpus (ex. le corpus « *Phonologie du français contemporain* » (PFC), (Durand *et al.*, 2009 ; Nguyen et Adda-Decker, 2013 ; Detey *et al.*, 2016a)), les développements de la sociophonétique et de la dialectologie perceptive ont également révélé l'intérêt, sinon la nécessité, de coupler analyses perceptives et études acoustiques si l'on souhaite approcher la réalité cognitive des mécanismes de catégorisation phonético-phonologique par les locuteurs-auditeurs natifs. L'analyse acoustique (ou articulatoire) des productions ne suffit pas à rendre compte de la classification psycho-acoustique effectuée par les auditeurs, certains traits ou détails phonétiques étant perceptivement jugés saillants ou, au contraire, négligés, selon les locuteurs-auditeurs. Ce couplage méthodologique perception-production a ainsi été effectué dans certaines études en français pour la caractérisation des accents régionaux (Boula de Mareüil *et al.*, 2013), ainsi que, à titre exploratoire, pour le français de référence (Detey *et al.*, 2016a). La question centrale du modèle de référence ne peut donc se limiter à une question de moyennage acoustique, non seulement du point de vue de la généralisabilité linguistique des données sur lesquelles il s'appuie, mais surtout en raison des objectifs des apprenants qui ne se résument plus aujourd'hui à « parler comme un natif ». Outre la nécessité d'intégrer la variation dans le système de référence (pour différents systèmes phonémiques vocaliques de référence, voir Detey *et al.*, 2016a, 60), de manière à déboucher à terme sur des systèmes idéalement polylectaux (possibilité de sélectionner le système de

référence au système cible visé, variable selon les communautés), cette orientation, couplée aux résultats de tests perceptifs, conduirait également à une procédure d'évaluation sensible aux degrés d'acceptabilité sociolinguistique des productions, si possible en termes d'intelligibilité, de compréhensibilité et de précision, davantage axées sur des frontières floues que catégoriques. La question du rapport entre normes natives et non natives, tant en production qu'en perception, pour les apprenants se trouve donc au cœur de la réflexion, et il faut, pour ces dernières, y inclure les apports des corpus d'apprenants.

4.2. *De la prise en compte des interlangues : les corpus non natifs*

L'intérêt de l'intégration de connaissances sur la L1 des apprenants ainsi que sur leurs interlangues (et donc leurs « erreurs ») dans les systèmes de correction de la prononciation (humain ou automatisé) n'est aujourd'hui plus à démontrer (comme le montre l'analyse de corpus permettant d'aboutir à un inventaire robuste d'erreurs de prononciation les plus fréquentes, persistantes et pouvant entraver la communication, décrite dans Neri *et al.*, 2006). L'intérêt de développer des programmes spécifiques à une L1 particulière a été souligné par plusieurs chercheurs (Burgos *et al.*, 2013), posant des questions méthodologiquement cruciales concernant la transcription et l'annotation (Carranza *et al.*, 2014 ; Carranza, 2016). Si l'intérêt de la transcription (orthographique ou phonologique) et de l'alignement automatique est ici évident (Strik et Cucchiarini, 2014), la transcription de la parole non native, plus encore que celle de la parole native, pose en effet des difficultés et soulève des questions parfois escamotées par l'automatisme de certaines procédures, qu'il s'agisse de cas d'indécidabilité ou de désaccords inter-transcripteurs et inter-évaluateurs (Zechner *et al.*, 2009), accentués par le caractère intrinsèquement instable et perméable (à la L1, à la L2 et à d'autres structures émergentes) du système interphonologique des apprenants (Racine *et al.*, 2011). En effet, si la multiplication progressive des corpus oraux d'apprenants de différentes L1 et L2 ouvre la voie à des systèmes mieux informés (Trouvain *et al.*, 2015), et si l'on peut espérer obtenir à terme des systèmes automatiques d'évaluation de la parole non native (Zechner *et al.*, 2009), de réels défis méthodologiques restent posés, à commencer par celui du lien entre représentations visuelles du signal (transcriptions, annotations) et visées du corpus. Trois approches sont envisageables : (i) *aucune transcription* : comparaison des formes acoustiques ; (ii) *transcription phonologique ou phonétique à l'aide de systèmes tels que l'API, SAMPA, ARPABET, etc.* : comparaison des formes ainsi transcrites avec des formes de référence stockées dans des lexiques et utilisation de catégories telles « substitution », « effacement », « insertion », etc. Il s'agit d'une des approches les plus fréquemment adoptées dans le domaine (ex. le corpus AESOP et son exploitation, Kondo *et al.*, 2015) ; (iii) *transcription orthographique couplée à une approche auditive* : catégorisation psychoacoustique sous forme d'évaluation perceptive en termes de degré de conformité à la cible par exemple, qui s'inscrit à la fois dans le principe même de la phonologie de corpus, en particulier en L2 (ne pas précatégoriser des catégories précisément en cours de construction), et en même temps dans une orientation pédagogique destinée

à traiter en priorité les éléments perçus par les auditeurs natifs non experts, avant de se pencher sur le détail phonétique fin des productions (par exemple dans le projet Interphonologie du français contemporain (IPFC), voir Detey *et al.*, 2016b). La valeur ajoutée de l'évaluation humaine d'une part, et celle de l'intégration de formes non natives aux lexiques de référence d'autre part, de manière à tenir compte des normes non natives, sont à présent bien repérées dans le domaine (Detey, 2012), et l'usage de corpus d'apprenants pour améliorer la détection et la correction automatique d'erreurs d'apprenants occupe de nombreux chercheurs (Gamon *et al.*, 2013). Or, l'approche (iii) attire l'attention des concepteurs (et des utilisateurs) des systèmes d'EPAO sur la définition de la cible : quelle flexibilité faut-il se permettre dans l'établissement de la cible et à quel niveau (phonétique, phonologique) ? Quel type de mesure adopter pour l'évaluation de l'écart entre la production et la cible ? Sur le plan phonétique la cible doit-elle être définie en termes de catégorie phonémique, en termes de traits phonétiques, de mesure acoustique ou de catégorisation psycho-acoustique ? Enfin, elle souligne également la nécessité d'intégrer l'évaluation humaine non seulement experte mais également non experte dans la procédure d'évaluation. Outre cette dimension méthodologique, les apports des corpus d'apprenants se situent évidemment dans la description qu'ils doivent à terme offrir des stades et parcours développementaux typiques d'une population donnée (en distinguant les plans phonétique et phonologique, ainsi que perceptif et productif), de manière à élaborer des modèles phonético-phonologiques dynamiques, possiblement en lien, dans le domaine didactique, avec les différents niveaux et les descripteurs du « Cadre européen commun de référence pour les langues » (CECRL) par exemple (Conseil de l'Europe, 2001 ; Detey et Racine, 2012). Ces modèles pourraient également tenir compte des tâches impliquées (en particulier lecture *vs* non lecture), et aux différents stades pourraient être attribués des degrés d'acceptabilité plus ou moins grands, en fonction de la précision phonétique des productions, mais aussi de la charge fonctionnelle des structures en question, ainsi que de leur catégorisation perceptive. En effet, une autre limite de l'utilisation qui est faite des corpus d'apprenants pour le développement de systèmes d'EPAO réside dans le fait que ces données servent surtout à améliorer la détection et la caractérisation automatique d'erreurs (Chen et Jang, 2012) et qu'il n'est pas proposé aujourd'hui de véritable réflexion quant à la progression à suivre dans l'apprentissage. Si une simple perspective « *data-driven* » oriente les concepteurs vers les erreurs les plus fréquentes en priorité (Burgos *et al.*, 2013), le point de vue didactique consiste généralement à traiter d'abord les erreurs les moins difficiles (et donc aussi les moins fréquentes) pour faire ensuite progresser l'apprenant vers les sons et/ou les positions phonologiques qui lui posent le plus de difficultés.

4.3. *Perspectives didactiques : format et* feedback

Si les systèmes évoqués plus haut pourraient mieux répondre aux besoins et aux attentes des utilisateurs, il va de soi que leur succès dépendra également de leur format didactique et du *feedback* qu'ils renverront à l'apprenant. On peut déjà noter que si le système d'EPAO est trop « strict » et qu'il rejette massivement les productions

d'apprenants (détection exhaustive des erreurs pour un débutant par exemple), ou s'il lui fournit trop d'informations données simultanément, le résultat peut être contre-productif et conduire l'apprenant à abandonner le système. Par-delà les aspects génériques (ex. langue et métalangage employés), de nombreuses études ont mis en évidence l'intérêt de l'instruction explicite (Saito et Saito, *sous presse*), de l'entraînement audiovisuel (Hazan *et al.*, 2005) ou encore de l'entraînement à haut degré de variabilité (Thomson, 2011) pour aider les apprenants à améliorer leur perception ou leur production, tant sur le plan phonémique que prosodique. Des études récentes nous renseignent sur l'impact différentiel de certains formats pédagogiquement classiques de retour correctif, comme celle de Gooch, Saito et Lyster à propos du « *recast* » et du « *prompt* » dans l'apprentissage de la liquide /ɹ/ par des apprenants coréens d'anglais recevant un enseignement centré sur la forme et orienté vers le sens (« *simulated meaning-oriented classrooms receiving form-focused instruction* ») (2016) : « *students were pushed by prompts to improve intelligibility mainly through the adjustment of interlanguage strategies (e.g., prolonging the phonemic length), and by recasts to refine accuracy in their /ɹ/ production* ». Ces études nous apprennent également que le retour correctif portant sur les erreurs de perception peut aider à améliorer la précision de la production, mais que cet effet dépend du type de retour correctif (Lee et Lyster, 2016). L'intérêt de la dimension visuelle, quant à lui, a déjà particulièrement été exploré (Hardison, 2007), au niveau suprasegmental (à travers la visualisation de courbes sonores (Cazade, 1999), mais aussi avec d'autres formats tels que des flashs lumineux, voir Hincks et Edlund, 2009), et, plus récemment, sur le plan segmental (voir Olson (2014) ; pour une revue des travaux et la proposition d'un paradigme de *feedback* visuel, ainsi que Offerman et Olson, 2016). Sans pouvoir couvrir ici l'ensemble des études concernant le retour correctif sur la production orale des apprenants de langue étrangère (pour une revue récente voir Brown, 2016), il est sans doute utile de rappeler les six types de retours identifiés par Lyster et Ranta (1997) dans leur étude phare, à savoir : « *recast* » (une manière de corriger une erreur implicitement sans bloquer la communication, essentiellement en répétant la forme produite par l'apprenant en la corrigeant, c'est-à-dire une reformulation corrective de la production erronée)[9] ; *correction explicite* ; *élicitation* (obtenir la forme correcte par l'apprenant plutôt que lui donner) ; *requête de clarification* ; *retour métalinguistique* ; *répétition de l'erreur*. Les auteurs soulignent que l'une des faiblesses du « *recast* », majoritairement employé par les enseignants parmi les six types identifiés, est qu'il entraîne souvent une ambiguïté sur le focus de l'intervention de l'enseignant (sur la forme ou sur le sens ?), ambiguïté éliminée par les autres types de retours correctifs ; l'engagement des apprenants dans le processus correctif semble le plus productif lorsque la forme correcte n'est pas fournie directement (comme avec le « *recast* » ou la correction explicite) (Lyster et Ranta 1997, p. 57-58). À la lecture de qui précède, il apparaît que l'évaluation quantifiée (taux de réussite ou autre score comparable) à elle seule ne suffit pas à aider l'apprenant à corriger sa prononciation : le lien entre lexique et phonolo-

9. Lyster et Ranta indiquent que le terme « écho » est parfois utilisé pour traduire « *recast* » en français, car il arrive que les apprenants ne perçoivent pas la différence entre leur production originale et la forme corrigée produite par leur enseignant (1997, p. 57).

gie, notamment, est fléché tant dans les travaux relatifs aux connexions entre forme et sens dans l'acquisition d'une L2 (Isaacs, 2009) que dans certains des modèles les plus récents en interphonologie (van Leussen et Escudero, 2015), tandis que différentes techniques facilitant l'acquisition ont été répertoriées. Il reste donc à intégrer ces derniers acquis aux procédures de remédiation offerts par les systèmes d'EPAO à ses utilisateurs.

5. Conclusion

Cet article vise un double lectorat : d'une part les ingénieurs de la parole, afin de les sensibiliser aux pratiques et aux perspectives contemporaines en didactique des langues, en particulier vis-à-vis des objectifs de la correction de la prononciation (intelligibilité *vs* précision phonétique), en lien avec l'élargissement des modèles de référence (variation native et non native), mais aussi vis-à-vis des attentes en termes de *feedback* (dépassement du niveau segmental et lien avec les études en didactique sur les effets de l'instruction, de la modalité et des retours correctifs) ; d'autre part les didacticiens pour leur présenter l'intérêt des systèmes d'EPAO et de leurs développements les plus récents. La première partie de l'article offre un aperçu des pratiques et des enjeux de l'enseignement/apprentissage de la prononciation en langue étrangère tel qu'il est effectué sans système automatisé, tandis que la deuxième partie décrit les procédures de reconnaissance automatique de la parole et leurs principales applications dans le domaine de la correction automatique de la prononciation. La troisième partie, enfin, se penche sur les pistes à explorer si l'on souhaite améliorer les interactions entre les deux champs, à savoir celle, d'obédience sociolinguistique, des modèles de référence pour l'évaluation de la prononciation, celle, d'obédience psycholinguistique, de l'intégration des connaissances des interlangues des apprenants pour l'optimisation des systèmes, et enfin celle, d'obédience didactique, du format et du contenu des retours correctifs que doivent fournir les systèmes d'EPAO en tenant compte des deux premières pistes.

Parmi les points de discussion pour le futur figurent ainsi notamment : (i) le rapport entre les modèles de langage statistiques construits à partir de grands corpus de textes pour les systèmes de RAP et la nature des corpus en question (oral *vs* écrit, type de discours, variétés de référence, etc.), au regard des avancées récentes effectuées en linguistique de corpus ; (ii) le rapport entre les seuils employés dans les systèmes de RAP pour décider de la qualité d'une réalisation phonétique et les seuils d'acceptabilité envisagés par les didacticiens travaillant en sociophonétique de corpus, et traitant des variétés natives et non natives (Galazzi et Guimbretière, 1991 ; Detey et Racine, 2012) ; (iii) le traitement de la parole spontanée et les cas d'ambiguïtés morpho-phonologiques (par opposition à de la parole lue ou répétée pour laquelle les cibles sont prédéfinies).

Concernant les systèmes complets d'EPAO, trois chantiers doivent être mentionnés : (1) *un manque d'études longitudinales à long terme* portant sur des systèmes d'EPAO complets, allant de la détection d'erreurs et du diagnostic à la remédiation

(*feedback*), puisque les progrès sont longs à accomplir et qu'ils peuvent se situer à une autre échelle que celles généralement adoptées dans les études psycholinguistiques ponctuelles ; (2) *un manque de prise en compte de la variation inter- et intra-apprenants* : l'expérience enseignante révèle que, même pour une L1 et un niveau communs, les profils d'apprenants sont très divers, en termes de difficultés de prononciation mais aussi et surtout en termes d'efficacité des types d'exercices et de *feedback* utilisés. De même, cette variabilité peut apparaître sur le plan intra-individuel, un apprenant ne bénéficiant pas au même degré d'un même type d'exercices ou de *feedback* au cours de son parcours d'apprentissage, d'où la nécessité de faire évoluer le travail de « profilage » automatique des apprenants pour intégrer cette variabilité et optimiser les exercices proposés ; (3) *un manque d'intégration de la prosodie dans les premières étapes de l'apprentissage* : du côté de l'ingénierie de la parole, la prosodie est envisagée comme un élément à travailler chez les apprenants les plus avancés, apportant plus de naturel à une parole déjà intelligible (Ai, 2013 ; Witt, 2012). Or, en didactique, le point de vue est généralement opposé, la prosodie étant considérée comme primordiale (première dans l'acquisition de la L1, support pour l'apprentissage segmental qui va suivre, notamment dans la MVT, etc.). Il reste donc à mener des études permettant de tester, dans des contextes didactiques authentiques et pour des langues sources et cibles spécifiques, des systèmes de RAP intégrant plus avant la variation en L1 ainsi que les variantes en L2, et ce, non seulement sur le plan segmental, mais aussi prosodique et multimodal, et offrant si possible des grilles d'évaluation adossables aux descripteurs des compétences en langue définis par exemple par le CECRL. Les chantiers à venir ne se situent donc plus seulement sur le plan de l'ingénierie de la parole, mais sur celui de la collaboration interdisciplinaire entre ingénieurs, didacticiens et enseignants.

Remerciements

Les réflexions présentées dans cet article ont bénéficié du soutien de JSPS KAKENHI JP 15H03227 et JP 23320121. Nous remercions trois évaluateurs anonymes pour leurs commentaires, ainsi que les membres des projets PFC et IPFC, en particulier Jacques Durand, Bernard Laks, Chantal Lyche, Isabelle Racine et Yuji Kawaguchi.

6. Bibliographie

Abdel-Hamid O., Mohamed A.-R., Jiang H., Penn G., « Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition », *Proc. ICASSP*, p. 4277-4280, 2012.

Ai R., « Perceptual Feedback in Computer Assisted Pronunciation Training : A Survey. », *RANLP*, p. 1-6, 2013.

Akahane-Yamada R., Tohkura Y., Bradlow A. R., Pisoni D. B., « Does training in speech perception modify speech production ? », *Proc. ICSLP*, p. 606-609, 1996.

Amodei D. *et al.*, « Deep speech 2 : End-to-end speech recognition in English and Mandarin », *Proc. ICML*, 2015.

Black M. *et al.*, « Automated Evaluation of Non-Native English Pronunciation Quality : Combining Knowledge- and Data-Driven Features at Multiple Time Scales », *Proc. Interspeech*, Dresde, p. 493-497, 2015.

Bohn O.-S., Munro M. (eds), *Language Experience in Second Language Speech Learning*, John Benjamins, 2007.

Boula de Mareüil P., Woehrling C., Adda-Decker M., « Contribution of automatic speech processing to the study of Northern/Southern French », *Lang Sci*, vol. 39, p. 75-82, 2013.

Brown A., « Functional Load and the teaching of pronunciation », *Tesol Quaterly*, vol. 22, n° 4, p. 593-606, 1988.

Brown D., « The type and linguistic foci of oral corrective feedback in the L2 classroom : A meta-analysis », *Lang Teach Res*, vol. 20, n° 4, p. 436-458, 2016.

Burgos P., Cucchiarini C., Van Hout R., Strik H., « Pronunciation errors by Spanish learners of Dutch : a data-driven study for ASR-based pronunciation training. », *Proc. Interspeech*, p. 2385-2389, 2013.

Calbris G., « La prononciation et la correction phonétique », *Le français dans le monde*, vol. 65, p. 28-37, 1969.

Calliope L., *Parole et son traitement automatique*, Masson Paris, 1989.

Carranza M., « Transcription and annotation of a Japanese accented spoken corpus of L2 Spanish for the development of CAPT applications », *in* A. Pareja-Lora, C. Calle-Martínez, P. Rodríguez-Arancón (eds), *New perspectives on teaching and working with languages in the digital era*, Research-publishing.net, p. 339-349, 2016.

Carranza M., Cucchiarini C., Burgos P., Strik H., « Non-native speech corpora for the development of computer assisted pronunciation training systems », *Proceedings of Edulearn 2014*, IATED, Valence, p. 3624-3633, 2014.

Cazade A., « De l'usage des courbes sonores et autres supports graphiques pour aider l'apprenant en langues », *Alsic*, vol. 2, n° 2, p. 3-32, 1999.

Champagne-Muzar C., Bourdages J. S., *Le point sur la phonétique*, Clé International, 1998.

Chen L.-Y., Jang J.-S. R., « Improvement in Automatic Pronunciation Scoring using Additional Basic Scores and Learning to Rank », *Proc. Interspeech*, Portland, p. 1295-1298, 2012.

Chun D., « Signal analysis software for teaching discourse Intonation », *Lang Learn Technol*, vol. 2, n° 1, p. 61-77, 1998.

Colantoni L., Steele J., Escudero P., *Second Language Speech. Theory and Practice*, Cambridge University Press, 2015.

Conseil de l'Europe, *Cadre européen commun de référence pour les langues*, Paris : Didier, 2001.

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

Detey S., « Coding an L2 phonological corpus : from perceptual assessment to non-native speech models – an illustration with French nasal vowels », *in* Y. Tono, Y. Kawaguchi, M. Minegishi (eds), *Developmental and crosslinguistic perspectives in learner corpus research*, John Benjamins, Amsterdam/Philadelphie, p. 229-250, 2012.

Detey S., « Enseignement de la prononciation et correction phonétique : principes essentiels », *in* S. Detey, I. Racine, Y. Kawaguchi, J. Eychenne (eds), *La prononciation du français dans le monde : du natif à l'apprenant*, CLE International, Paris, p. 226-235, 2016.

Detey S., Lyche C., Racine I., Schwab S., Gac D. L., « The notion of norm in spoken French : production and perception », *in* S. Detey, J. Durand, B. Laks, C. Lyche (eds), *Varieties of Spoken French*, Oxford University Press, p. 55-67, 2016a.

Detey S., Racine I., « Les apprenants de français face aux normes de prononciation : quelle(s) entrée(s) pour quelle(s) sortie(s) ? », *Revue française de linguistique appliquée*, vol. 17, n° 1, p. 81-96, 2012.

Detey S., Racine I., Kawaguchi Y., Zay F., « Variation among non-native speakers : the Inter-Phonology of Contemporary French », *in* S. Detey, J. Durand, B. Laks, C. Lyche (eds), *Varieties of Spoken French*, Oxford University Press, p. 489-502, 2016b.

Durand J., Gut U., Kristoffersen G., *The Oxford Handbook of Corpus Phonology*, Oxford University Press, 2014.

Durand J., Laks B., Lyche C., « Le projet PFC : une source de données primaires structurées », *in* J. Durand, B. Laks, C. Lyche (eds), *Phonologie, variation et accents du français*, Hermès, p. 19-61, 2009.

Edwards J. G. H., Zampini M. L. (eds), *Phonology and Second Language Acquisition*, John Benjamins, 2008.

Ehsani F., Knodt E., « Speech technology in computer-aided language learning », *Lang Learn Technol*, vol. 1, n° 2, p. 45-60, 1998.

Eskenazi M., « Using automatic speech processing for foreign language pronunciation tutoring : some issues and a prototype », *Lang Learn Technol*, vol. 2, n° 2, p. 62-76, 1999.

Eskenazi M., « An overview of spoken language technology for education », *Speech Commun*, vol. 51, p. 832-844, 2009.

Fonix Talk SDK, *Speech FX Text to Speech*, 2016. Online : http ://www.speechfxinc.com.

Fontan L., Tardieu J., Gaillard P., Woisard W., Ruiz R., « Relationship Between Speech Intelligibility and Speech Comprehension in Babble Noise », *J speech lang hear r*, vol. 58, p. 977-986, 2015.

Galazzi E., *Le son à l'école. Phonétique et enseignement des langues (fin XIXe siècle - début XXe siècle)*, La Scuola, 2002.

Galazzi E., Guimbretière E., « Seuil d'acceptabilité des réalisations prosodiques d'apprenants italophones », *Atti del 2 Convegno Internazionale di analisi comparativa francese/italiano*, vol. 2, p. 104-120, 1991.

Gamon M., Chodorow M., Leacock C., Tetreault J., « Using learner corpora for automatic error detection and correction », *in* A. Diaz-Negrillo, N. Ballier, P. Thompson (eds), *Automatic treatment and analysis of learner corpus data*, John Benjamins, p. 127-149, 2013.

Gendrot C., Adda-Decker M., « Impact of duration on F1/F2 formant values of oral vowels : an automatic analysis of large broadcast news corpora in French and German. », *Variations*, vol. 2, n° 22.5, p. 2-4, 2005.

Georgeton L., Paillereau N., Landron S., Gao J., Kamiyama T., « Analyse formantique des voyelles orales du français en contexte isolé : à la recherche d'une référence pour les apprenants de FLE », *Conférence JEP-TALN-RECITAL*, p. 145-152, 2012.

Gooch R., Saito K., Lyster R., « Effects of recasts and prompts on L2 pronunciation development : Teaching English /r/ to Korean adult EFL learners », *System*, vol. 60, p. 117-127, 2016.

Guberina P., « Les appareils Suvag et Suvag Lingua », *Revue de Phonétique Appliquée*, vol. 27-28, p. 7-16, 1973.

Guberina P., Gospodnetic N., Pozojenic M., Skaaric P., Vuletic B., « Correction de la prononciation des élèves qui apprennent le français », *Revue de Phonétique Appliquée*, vol. 1, p. 81-94, 1965.

Hansen T. K., « Computer assisted pronunciation training : the four 'k's of feedback », *Proc. m-ICTE*, Séville, p. 342-346, 2006.

Hardison D. M., « The visual element in phonological perception and learning », *in* M. C. Pennington (ed.), *Phonology in context*, Palgrave Macmillan, New York, p. 135-158, 2007.

Hazan V., Sennema A., Iba M., Faulkner A., « Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English », *Speech commun*, vol. 47, n° 3, p. 360-378, 2005.

Hincks R., Edlund J., « Promoting increased pitch variation in oral presentations with transient visual feedback », *Lang Learn Technol*, vol. 13, n° 3, p. 32-50, 2009.

Hu W., Qian Y., Soong F., « A New DNN-based High Quality Pronunciation Evaluation for Computer-Aided Language Learning (CALL) », *Proc. Interspeech*, Lyon, p. 1886-1890, 2013.

Hu W., Qian Y., Soong F., Wang Y., « Improved Mispronunciation Detection With Deep Neural Network Trained Acoustic Models and Transfer Learning based Logistic Regression Classifiers », *Speech Commun*, vol. 67, p. 154-166, 2015.

Hueber T., « Ultraspeech-player : Intuitive visualization of ultrasound articulatory data for speech therapy and pronunciation training », *Proc. Interspeech*, Lyon, p. 752-753, 2013.

Huensch A., « Perceptual phonetic training improves production in larger discourse contexts », *Journal of Second Language Pronunciation*, vol. 2, n° 2, p. 183-207, 2016.

Intravaia P., *Formation des professeurs de langue en phonétique corrective. Le système verbotonal*, Didier Erudition, 2000.

Isaacs T., « Integrating form and meaning in L2 pronunciation instruction », *TESL Canada Journal*, vol. 27, n° 1, p. 1-12, 2009.

Isaacs T., « Assessing speaking », *in* D. Tsagari, J. Banerjee (eds), *Handbook of Second Language assessment*, DeGruyter Mouton, Berlin, p. 131-146, 2016.

Isaacs T., Thomson R. I., « Rater experience, rating scale length, and judgments of L2 pronunciation : revisiting research conventions », *Language Assessment Quarterly*, vol. 10, n° 2, p. 135-159, 2013.

Kang O., Moran M., « Functional loads of pronunciation features in nonnative speakers' oral assessment », *Tesol Quarterly*, vol. 48, n° 1, p. 176-187, 2014.

Kondo M., Tsubaki H., Sagisaka Y., « Segmental variation of Japanese speakers' English : Analysis of "the North Wind and the Sun" in AESOP corpus », *Journal of the Phonetic Society of Japan*, vol. 19, p. 3-17, 2015.

Laborde V., Pellegrini T., Fontan L., Mauclair J., Sahraoui H., Farinas J., « Pronunciation Assessment of Japanese Learners of French with GOP Scores and Phonetic Information », *Proc. Interspeech*, San Francisco, p. 2686-2690, 2016.

Laks B., « Description de l'oral et variation : la phonologie et la norme », *Information Grammaticale*, vol. 94, p. 5-11, 2002.

Lauret B., *Enseigner la prononciation du français : questions et outils*, Hachette, 2007.

Lee A., Lyster R., « Can corrective feedback on second language speech perception errors affect production accuracy ? », *Appl psycholinguist*, 2016.

Li W., Li K., Siniscalchi S. M., Chen N. F., Lee C.-H., « Detecting Mispronunciations of L2 Learners and Providing Corrective Feedback Using Knowledge-guided and Data-driven Decision Trees », *Proc. Interspeech*, San Francisco, p. 3127-3131, 2016.

Lu J., Wang R., Silva L. C., « Automatic stress exaggeration by prosody modification to assist language learners perceive sentence stress », *International Journal of Speech Technology*, vol. 15, n° 2, p. 87-98, 2002.

Lu Y., *Etude contrastive de la prosodie audio-visuelle des affects sociaux en chinois mandarin vs. français : vers une application pour l'apprentissage de la langue étrangère ou seconde*, Université Grenoble Alpes, 2015. Thèse de doctorat.

Luo D., Qiao Y., Minematsu N., Yamauchi Y., Hirose K., « Analysis and utilization of MLLR speaker adaptation technique for learners' pronunciation evaluation », *Proc. Interspeech*, Brigthon, p. 608-611, 2009.

Lyche C., « Le français de référence : éléments de synthèse », *in* S. Detey, J. Durand, B. Laks, C. Lyche (eds), *Les variétés du français parlé dans l'espace francophone : ressources pour l'enseignement*, Ophrys, Paris, 2010.

Lyster R., Ranta L., « Corrective feed-back and learner uptake : Negotiation of form in communicative classrooms », *Stud Second Lang Acquis*, vol. 20, p. 37-66, 1997.

Martin P., « WinPitch LTL II, a multimodal pronunciation software », *Proc. InSTIL/ICALL*, Venise, 2004.

Meinedo H., Caseiro D., Neto J., Trancoso I., « AUDIMUS. media : a Broadcast News speech recognition system for the European Portuguese language », *International Workshop on Computational Processing of the Portuguese Language*, Springer, p. 9-17, 2003.

Meng F., Wu Z., Meng H., Jia J., Cai L., « Generating emphasis from neutral speech using hierarchical perturbation model by decision tree and support vector machine », *Proc. ICALP*, Warwik, 2012.

Miyakoda H., « Development of a pronunciation training system based on auditory-visual elements », *Proc. Interspeech*, Lyon, p. 2660-2661, 2013.

Montacié C., Caraty M.-J., « Phrase Accentuation Verification and Phonetic Variation Measurement for the Degree of Nativeness Sub-Challenge », *Proc. Interspeech*, Dresde, p. 483-487, 2015.

Morin Y.-C., « Le français de référence et les normes de prononciation », *Cahiers de l'Institut de linguistique de Louvain*, vol. 26, n° 1, p. 91-135, 2000.

Moulton W. G., « Toward a classification of pronunciation errors », *Mod Lang J*, vol. 40, n° 3, p. 101-109, 1962.

Moyer A., *Foreign Accent. The Phenomenon of Non-native Speech*, Cambridge University Press, 2013.

Munro M. J., Derwing T. M., « Foreign accent, comprehensibility, and intelligibility in the speech of second language learners », *Lang Learn*, vol. 45, n° 1, p. 73-97, 1995.

Munro M. J., Derwing T. M., « The functional load principle in ESL pronunciation instruction : An exploratory study », *System*, vol. 34, nº 4, p. 520-531, 2006.

Neri A., Cucchiarini C., Strik H., « Feedback in Computer Assisted Pronunciation Training : Technology push or demand pull ? », *Proc. ICSLP*, Denver, p. 1209-1212, 2002.

Neri A., Cucchiarini C., Strik H., « Selecting segmental errors in non-native Dutch for optimal pronunciation training », *IRAL*, vol. 44, nº 4, p. 357-404, 2006.

Nguyen N., Adda-Decker M. (eds), *Méthodes et outils pour l'analyse phonétique des grands corpus oraux*, Hermes Science Publications, 2013.

Offerman H. M., Olson D. J., « Visual feedback and second language segmental production : The generalizability of pronunciation gains », *System*, vol. 59, p. 45-60, 2016.

Olson D., « Benefits of visual feedback on segmental production in the L2 classroom », *Lang Learn Technol*, vol. 18, nº 3, p. 173-92, 2014.

Pellegrini T., Fontan L., Sahraoui H., « Réseau de neurones convolutif pour l'évaluation automatique de la prononciation », *Conférence JEP-TALN-RECITAL*, Paris, p. 624-632, 2016.

Pennington M. C., Richards J. C., « Pronunciation revisited », *Tesol Quaterly*, vol. 20, nº 2, p. 207-225, 1986.

Pennington M. C., Richards J. C., « Computer-aided pronunciation pedagogy : promise, limitations, directions », *Computer Assisted Language Learning*, vol. 12, nº 5, p. 427-440, 1999.

Qian Y., Wang X., Evanini K., Suendermann-Oeft D., « Self-Adaptive DNN for Improving Spoken Language Proficiency Assessment », *Proc. Interspeech*, San Francisco, p. 3122-3126, 2016.

Rabiner L. R., « A tutorial on hidden Markov models and selected applications in speech recognition », *Proceedings of the IEEE*, vol. 77, nº 2, p. 257-286, 1989.

Racine I., Zay F., Detey S., Kawaguchi Y., « De la transcription de corpus à l'analyse interphonologiques : enjeux méthodologiques en FLE », 2011.

Renard R. (ed.), *Apprentissage d'une langue étrangère/seconde 2. La phonétique verbo-tonale*, De Boeck Université, 2002.

Rilliard A., Shochi T., Martin J.-C., Erickson D., Aubergé V., « Multimodal indices to Japanese and French prosodically expressed social affects », *Lang speech*, vol. 52, nº 2/3, p. 223-243, 1988.

Rivenc P. (ed.), *Apprentissage d'une langue étrangère/seconde. Vol. 3 : la méthodologie*, De Boeck, 2003.

Saito K., « Effects of instruction on L2 pronunciation development : A synthesis of 15 quasi-experimental intervention studies », *TESOL Quarterly*, vol. 46, p. 842-854, 2012.

Saito Y., Saito K., « Differential effects of instruction on the development of second language comprehensibility, word stress, rhythm, and intonation : The case of inexperienced Japanese EFL learners », *Lang Teach Res*, *sous presse*.

Saz O., Eskenazi M., « Addressing Confusions in Spoken Language in ESL Pronunciation Tutors », *Proc. Interspeech*, Portland, p. 771-774, 2012.

Strik H., Cucchiarini C., *On automatic phonological transcription of speech corpora*, Oxford University Press, 2014.

Tao J., Chen L., Lee C. M., « DNN Online with iVectors Acoustic Modeling and Doc2Vec Distributed Representations for Improving Automated Speech Scoring », *Proc. Interspeech*, San Francisco, p. 3117-3121, 2016.

TELL ME MORE, *Language learning software*, 2013. Online : http ://www.tellmemore.com.

Thomson R. I., « Computer-Assisted Pronunciation Training : Targeting second language vowel perception improves pronunciation », *CALICO Journal*, vol. 28, nº 3, p. 744-765, 2011.

Thomson R. I., Derwing T. M., « The effectiveness of L2 pronunciation instruction : A narrative review », *Applied Linguistics*, vol. 36, p. 326-344, 2015.

Trofimovich P., Gatbonton E., « Repetition and focus on form in processing L2 Spanish words : Implications for pronunciation instruction », *Mod Lang J*, vol. 90, p. 519-535, 2006.

Troubetzkoy N. S., *Principes de phonologie*, Klincksieck, 1949.

Trouvain J., Zimmerer F., Gósy M., Bonneau A., « Phonetic Learner Corpora. Book of extended abstracts », *Satellite workshop of ICPhS2015*, Glasgow, 2015.

Vaissière J., Boula de Mareüil P., « Identifying a language or an accent : from segment to prosody », *Workshop MIDL*, Paris, 2004.

van Leussen J.-W., Escudero P., « Learning to perceive and recognize a second language : the L2LP model revised », *Frontiers in psychology*, 2015.

Vieru B., de Mareüil P. B., Adda-Decker M., « Identification and characterisation of non-native French accents », *Speech Commun*, vol. 53, p. 292-310, 2011.

Vogel K., *L'interlangue : la langue de l'apprenant*, Presses Universitaires du Mirail, 1995.

Wang Y.-B., Lee L.-S., « Error Pattern Detection Integrating Generative and Discriminative Learning for Computer-Aided Pronunciation Training », *Proc. Interspeech*, Portland, p. 819-822, 2012a.

Wang Y.-B., Lee L.-S., « Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training », *Proc. ICASSP*, Kyoto, p. 5049-5052, 2012b.

Wei W., Hu G., Hu Y., Wang R. H., « A new method for mispronunciation detection using support vector machine based on pronunciation space mode », *Speech Commun*, vol. 51, nº 10, p. 896-905, 2009.

Wilson E. O., Spaulding T. J., « Effects of noise and speech intelligibility on listener comprehension and processing time of Korean-accented English », *J speech lang hear r*, vol. 53, p. 1543-1554, 2010.

Witt S., *Use of Speech Recognition in Computer-Assisted Language Learning*, University of Cambridge, 1999. Thèse de doctorat.

Witt S., « 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, 2012.

Yuan Y., Liberman M., « Phoneme, Phone Boundary, and Tone in Automatic Scoring of Mandarin Proficiency », *Proc. Interspeech*, San Francisco, p. 2145-2149, 2016.

Zechner K., Higgins D., Xi X., Williamson D., « Automatic scoring of non-native spontaneous speech in tests of spoken English », *Speech Commun*, vol. 51, nº 10, p. 883-895, 2009.