Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Erreurs de prononciation en L2 : comparaison de méthodes pour la détection et le diagnostic guidés par la didactique

Romain Contrain[1] Julien Pinquier[1] Lionel Fontan[2] Isabelle Ferrané[1]
(1) IRIT, 118 Route de Narbonne, F-31062 Toulouse CEDEX 9, France
(2) Archean Labs, 20 place Prax-Paris, 82000 Montauban, France
romain.contrain@irit.fr, julien.pinquier@irit.fr, lfontan@archean.tech, isabelle.ferrane@irit.fr

RÉSUMÉ

La détection et diagnostic d'erreurs de prononciation nécessite des systèmes adaptés aux spécificités de la parole non-native. Élaborer de tels systèmes reste difficile à cause de la rareté des corpus dédiés incluant des annotations expertes. Dans cet article, nous proposons et comparons deux approches, l'une basée sur une transcription phonétique et l'autre sur l'alignement de signaux audio, élaborées dans le but de servir dans un programme d'entraînement à la prononciation assisté par ordinateur (EPAO). Nous les évaluons sur un corpus de parole non-native annoté selon des considérations didactiques, et nous trouvons que l'approche basée sur l'alignement a des propriétés préférables pour l'EPAO, dépassant la précision de l'autre approche de 31,1 % et 3,8 % en absolu sur deux erreurs communes des apprenants japonais du français.

ABSTRACT

L2 mispronunciations : a comparison of didactics-guided detection and diagnosis methods

Mispronunciation detection and diagnosis requires systems that are adapted to the specificities of non-native speech. Developing such models remains challenging due to the scarcity of non-native speech corpora and expert annotations. In this work, we propose and compare two approaches, one based on phonetic transcription and the other based on audio-to-audio alignment, meant to be used in computer-assisted pronunciation training (CAPT) software. We evaluate them on a corpus of non-native speech that was annotated following didactic considerations, and find that the alignment-based approach has preferable properties for CAPT, surpassing the precision of the other approach by 31.1 % and 3.8 % absolute on two common mispronunciations of Japanese learners of French.

MOTS-CLÉS : Entraînement à la Prononciation Assisté par Ordinateur, détection et diagnostic d'erreurs de prononciation, parole non-native, apprentissage profond.

KEYWORDS: Computer-Assisted Pronunciation Training, Mispronunciation Detection and Diagnosis, non-native speech, deep learning.

1 Introduction

La majorité des apprenants n'ayant pas accès à un professeur particulier pour travailler leur compétence orale, les outils d'entraînement à la prononciation assisté par ordinateur (EPAO) offrent alors un support pédagogique intéressant. Il est nécessaire que ces outils soient capables de détecter et de diagnostiquer les erreurs de prononciation avec suffisamment de fiabilité et de précision pour pouvoir

fournir à l'apprenant des retours pertinents vis-à-vis des difficultés qu'il rencontre.

Dans ce domaine, nombre de travaux s'appuient sur un alignement forcé du signal de parole avec la prononciation canonique, ce qui permet d'évaluer les sons produits en connaissant les phones canoniques auxquels ils correspondent. Dans cette approche, la méthode la plus utilisé est le GOP (Witt & Young, 2000) employé par exemple par Laborde *et al.* (2016) qui utilisent des scores issus de la variante F-GOP (combiné à d'autres informations) dans une régression logistique. Des méthodes plus récentes emploient des classifieurs basés sur des réseaux de neurones profonds (DNN) ou des représentations comme wav2vec 2.0 (Baevski *et al.*, 2020). Dans Sancinetti *et al.* (2022), un DNN utilisé dans un pipeline GOP est *fine-tuné* pour produire directement des probabilités de mauvaise prononciation. Pour détecter les erreurs de prononciation, Xu *et al.* (2021) se basent sur des frontières obtenues par un alignement forcé avec l'énoncé cible et sur des représentations issues de wav2vec 2.0 fournies à un réseau convolutionnel (CNN). Ces travaux se limitent cependant à de la détection d'erreur sans fournir de diagnostic.

D'autres méthodes se basent sur une transcription phonétique suivie d'une comparaison avec la prononciation canonique, ce qui permet de fournir un diagnostic d'erreur. La méthode de comparaison peut être un alignement de Needleman-Wunsch (Needleman & Wunsch, 1970), comme dans Leung *et al.* (2019), mais n'est pas toujours spécifiée. Lin & Wang (2022) entraînent conjointement un modèle sur la détection d'erreurs de prononciation et la reconnaissance de phones (apprentissage multi-tâches) pour produire des transcriptions phonétiques associées à des probabilités de mauvaise prononciation. Bien que les auteurs choisissent de se concentrer sur la détection, une telle approche pourrait servir également au diagnostic. Wu *et al.* (2021) expérimentent avec deux architectures basées sur des *transformers* pour effectuer la phase de reconnaissance de phones, une utilisant comme fonction de coût l'entropie croisée et une autre utilisant le coût CTC (Connectionist Temporal Classification).

Les travaux décrits précédemment reposent sur des modèles phonétiques appris sur de la parole non-native, parfois en relativement grande quantité (environ 30 heures pour Wu *et al.* (2021)). Cependant, les corpus de parole non-native dédiés à la détection et au diagnostic d'erreurs restent rares et peu volumineux par rapport aux corpus de parole native. Ceci est dû à la rareté relative des apprenants de langue et de l'expertise nécessaire pour annoter cette parole au niveau phonétique. Récemment, Korzekwa *et al.* (2022) propose de palier ce manque de données en générant de nouveaux exemples de mauvaises prononciations à partir d'exemples existants, tandis que Xu *et al.* (2021) apprend les caractéristiques de la parole non-native de manière auto-supervisée sur des données non annotées avant d'entraîner son modèle de détection d'erreurs sur une quantité plus faible de données annotées.

Les erreurs faites par les non-natifs sont souvent dues à l'influence de leur langue maternelle (L1) sur la langue apprise (L2) (Detey & Racine, 2016). Par conséquent, certains travaux se concentrent sur une paire L1/L2 donnée comme Sancinetti *et al.* (2022) et Laborde *et al.* (2016), ce dernier se concentrant sur la paire japonais/français, spécifiquement sur les phonèmes /ʁ/ et /v/, difficiles pour les japonophones (Kamiyama *et al.*, 2016). Certains travaux incorporent même des connaissances sur les erreurs communes pour cette paire, comme Ghosh *et al.* (2017) ou Harrison *et al.* (2009) qui utilise un *extended recognition network* qui modélise les schémas d'erreur probables pour la paire cantonnais/anglais.

Dans cet article, nous présentons deux approches de détection et diagnostic d'erreurs de prononciation appliquées à la paire japonais/français. Après avoir décrit le corpus de productions orales d'apprenants japonophones du français à notre disposition, nous présentons les deux approches développées. L'une

se base sur un système de transcription phonétique, tandis que l'autre s'appuie sur une méthode d'alignement entre signaux pour s'affranchir de l'apprentissage de modèles phonétiques. Nous les évaluons sur un sous-ensemble de phonèmes et d'erreurs cibles dont le choix a été guidé par des connaissances didactiques spécifiques à la paire de langues considérée. Enfin, nous comparons et discutons les résultats de chaque approche dans la perspective de leur intégration dans un système d'EPAO proposant une tâche de répétition de stimuli.

2 Corpus d'apprenants et catégories didactiques

Dans le cadre du laboratoire commun ALAIA, un corpus regroupant des productions d'étudiants japonais apprenant le français a été constitué (ci-après APPR). Il a été collecté auprès de 67 apprenants dans le cadre de tâches de répétition. Chaque apprenant devait répéter plusieurs stimuli parmi les 199 possibles, les stimuli présentés ayant été prononcés par le même locuteur français natif. Chaque stimulus est composé d'un seul mot ou d'une courte phrase (1 à 6 syllabes) choisis pour faire travailler certains sons de la langue apprise. Ce corpus de français L2 totalise 7112 enregistrements.

Un expert en interphonologie japonais/français a transcrit phonétiquement chaque production et mis en correspondance chaque phone produit avec le phonème attendu. Un second expert a ensuite annoté les mêmes productions et aligné temporellement les transcriptions phonétiques avec le signal audio correspondant. Nous disposons ainsi des segments correspondant à la réalisation de 47541 phonèmes, dont 20 % sont des erreurs de prononciation. Les annotateurs sont d'accord sur 82 % des énoncés et sur 96,7 % des phonèmes. Plusieurs réalisations ont pu être identifiées comme ambiguës (environ 4500 réalisations touchant 3639 enregistrements), parce qu'un annotateur a hésité ou parce que les annotateurs sont en désaccord (par exemple sur un phone qui serait entre [y] et [ɯ]). En retirant ces réalisations ambiguës, il reste environ 43000 réalisations dont 17 % sont des erreurs de prononciation.

Les réalisations ont été regroupées suivant une approche guidée par la didactique. Chaque catégorie didactique, définie du point de vue de la perception d'un locuteur natif, correspond à un type de difficulté rencontrée par les apprenants qui appelle à un type d'exercice de remédiation en particulier. Certaines réalisations sont trop peu nombreuses et trop atypiques pour justifier la création d'une catégorie propre et sont considérées comme « autres ». Le français compte environ une douzaine de phonèmes difficiles pour les japonophones, source d'erreurs fréquentes et importantes pour leur intelligibilité (Kamiyama *et al.*, 2016), et qui sont donc ceux qui nous intéressent pour notre étude. Le travail de définition de ces catégories, mené par un expert de l'enseignement du FLE, a été réalisé pour deux phonèmes d'intérêt : la voyelle /y/ et la consonne /ʒ/. La table 1 présente les catégories définies pour ces deux phonèmes. Le corpus compte respectivement 1182 et 1540 réalisations non-ambiguës de/y/ et /ʒ/. Les catégories sont déséquilibrées en terme d'effectifs, avec par exemple 74 % des réalisations de /y/ qui sont correctes, contre seulement 13 % de « perçu comme [ø;œ] ».

3 Systèmes de détection et diagnostic d'erreurs de prononciation

Nous comparons deux approches : l'une repose sur une transcription phonétique de la production de l'apprenant, et l'autre sur un alignement des segments audio correspondant à la prononciation attendue et à sa réalisation. La production de l'apprenant est analysée en connaissant la transcription phonétique du stimulus à répéter (prononciation canonique), et le phone cible, celui sur lequel nous

Catégorie	Exemples de réalisations	Effectif
/y/ correct	[y], [yh]	869
/y/ perçu comme [ø;œ]	[ø], [œ], [ɯ], [ə], [əh]	151
/y/ perçu comme [j+Voyelle]	[jy], [jɯ], [ju], [jə], [je], [jø]	87
autre réalisation de /y/	[u], [a], [o], [ɔ], [i], délétion	75
/ʒ/ correct	[ʒ]	647
/ʒ/ perçu comme [dʒ]	[dʒ], [tʃ]	862
/ʒ/ perçu comme [ʃ]	[ʃ]	19
autre réalisation de /ʒ/	[d], [t]	12

TABLE 1 – Classement des réalisations en catégories didactiques

cherchons une erreur de prononciation. Le système prédit la catégorie didactique de la réalisation, qui peut être « correct » ou une des catégories d'erreurs spécifiques au phonème (voir table 1).

3.1 Approche basée transcription

Dans l'approche basée sur la transcription phonétique, un modèle de reconnaissance de phones fournit une transcription phonétique de la production de l'apprenant. Ensuite, un alignement de Needleman-Wunsch avec la prononciation canonique permet d'identifier la partie de la transcription qui correspond au phone cible. Enfin, nous recherchons quelle catégorie didactique concorde le plus avec ce qui a été transcrit. Pour cela, nous calculons une mesure de similarité entre le(s) phone(s) transcrit(s) et des réalisations typiques de chaque catégorie. Les deux dernières étapes utilisent une matrice de similarité entre les phones basée sur la matrice de distances proposée par Ghio *et al.* (2018). La figure 1 donne un exemple du fonctionnement de ce système dans le cas d'une répétition de « je chante » où nous nous intéressons au phonème /ʒ/.

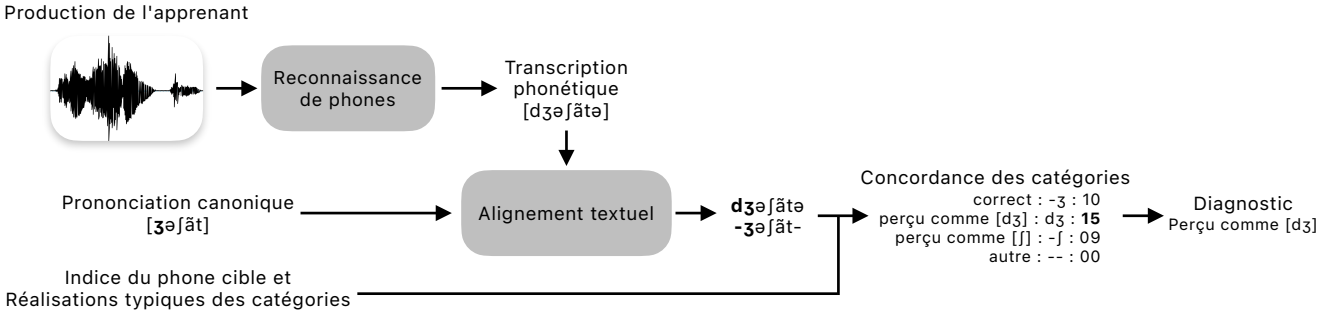

FIGURE 1 – Fonctionnement de l'approche basée sur la transcription phonétique

Le modèle de reconnaissance de phones employé se base sur Allosaurus [1], un modèle multilingue présenté par Li *et al.* (2020). Ce modèle utilise des unités BLSTM et une *loss* CTC. Il est conçu pour fonctionner avec n'importe quelle langue grâce à une étape de reconnaissance phonétique indépendante de la langue suivie d'une interprétation phonologique des phones transcrits. Cela permet de choisir n'importe quel jeu de phones pour la transcription. Pour l'inférence nous exploitons notre connaissance de l'énoncé à répéter pour restreindre le jeu de phones à ceux présents dans le mot et aux

1. https://github.com/xinjli/allosaurus v1.0

erreurs communes sur les phonèmes d'intérêt (s'ils sont présents). Par exemple, si la prononciation attendue est [ʒəʃãt], le jeu de phones sera restreint à [ʒ], [dʒ], [ʃ], [ə], [ã], [t].

Bien qu'il soit présenté comme universel, le modèle Allosaurus de base n'a jamais été entraîné sur du français et fonctionne mal sur cette langue : sur le corpus APPR il obtient un taux d'erreur phone (PER) d'environ 60 % (contre 25,0 % en moyenne sur ses langues d'entraînement) et il est notamment incapable de transcrire les voyelles nasales correctement. Nous avons donc dû pré-entraîner le modèle sur une quantité plus grande de parole native avant de l'adapter sur la petite quantité de parole non-native à notre disposition (voir section 2). Nous avons débuté avec un modèle monolingue pré-entraîné sur 150h de français natif issu du corpus Common Voice (ci-après Al-fr), puis nous avons expérimenté avec un modèle hybride pré-entraîné sur 150h de français et 65h de japonais natif issus du même corpus (ci-après Al-frjp). En effet, les apprenants ont tendance à produire des phones issus du système phonologique de leur L1 même s'ils sont absents de celui de la L2, et un modèle monolingue pourrait mal les gérer. Lors de l'adaptation à la parole non-native, chacun des modèles précédents a été entraîné sur le sous-ensemble des réalisations (issues d'APPR) correspondant aux phones qu'ils supportent (ces sous-ensembles sont nommés APPR-fr et APPR-frjp dans la section 4).

Common Voice (Ardila *et al.*, 2020) est un corpus multilingue de phrases lues, enregistrées par les contributeurs d'une plate-forme en ligne ouverte [2]. Pour le français, nous avons repris le sous-ensemble de 148,9 heures employé par Gelin *et al.* (2021) pour l'entraînement, et utilisé les 9,6 heures mises de côté par les auteurs comme ensemble de validation. Pour le japonais, 68,7 heures étaient exploitables dans la version 13.0. L'ensemble de validation a été construit pour représenter 6% du total en maximisant le nombre de locuteurs différents et en ayant le même ratio homme/femme que dans l'ensemble d'entraînement, aboutissant à 64,4 heures pour l'entraînement et 4,3 heures pour la validation. Les textes français ont été phonétisés à l'aide d'un dictionnaire de prononciation, et les japonais à l'aide de l'outil pykakasi [3] et de règles de prononciation.

3.2 Approche basée alignement

Un alignement temporel est réalisé entre le stimulus à répéter et de la production de l'apprenant au moyen de l'algorithme fastDTW (Salvador & Chan, 2007) et de représentations issues de wav2vec 2.0. Cela permet de faire correspondre la réalisation du phonème cible dans le stimulus, dont les frontières sont connues, avec sa réalisation dans la production. Ensuite, le segment correspondant est isolé, représenté avec un nouveau jeu de paramètres et des modèles de classification spécifiques au phonème cible déterminent la catégorie didactique de la réalisation. La figure 2 illustre ce principe de fonctionnement avec l'exemple d'une répétition de « le bus » où nous nous intéressons au phonème cible /y/.

Les deux étapes utilisent les mêmes représentations (wav2vec 2.0 [4]), mais la classification y ajoute des paramètres plus classiques, notamment 20 MFCC avec leurs dérivées premières et secondes, le Zero Crossing Rate et divers paramètres spectraux. Étant donné que la classification a besoin de données de dimension fixe, c'est la moyenne et l'écart-type sur le segment de chaque caractéristique qui sont fournis en entrée, ce qui donne une représentation à 252 dimensions.

La classification est réalisée par des *Random Forest* (Breiman, 2001) organisés en une architecture

2. https://commonvoice.mozilla.org
3. https://codeberg.org/miurahr/pykakasi
4. modèle pré-entraîné *wav2vec2-base-960h*

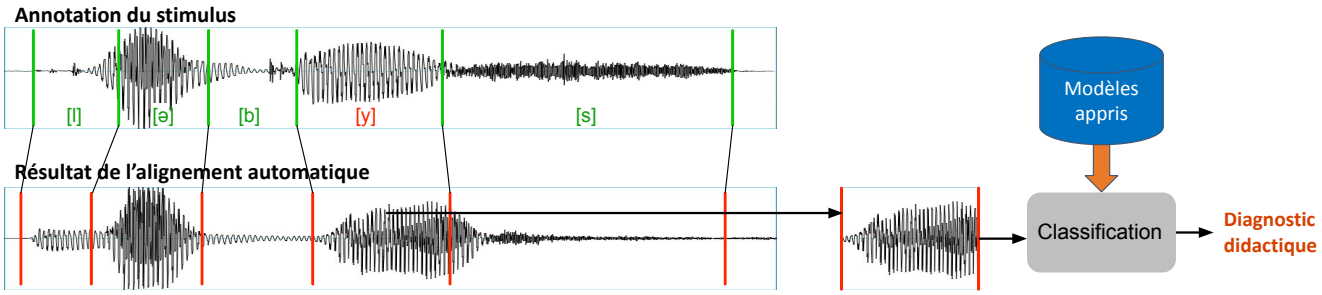

FIGURE 2 – Fonctionnement de l'approche basée alignement

hiérarchique binaire. Chacun est spécialisé dans l'appartenance à une catégorie didactique précise. Si le segment à classer n'appartient pas à la classe modélisée par le premier classifieur, alors nous regardons s'il appartient à la classe du deuxième, etc. jusqu'à atteindre le dernier niveau qui discrimine entre la dernière classe d'erreur et une classe « autre » utilisée comme une classe de rejet.

Les classifieurs pour chaque phonème d'intérêt ont été entraînés sur les réalisations de ce phonème dans le corpus APPR. Pour compenser le déséquilibre des classes, nous employons la technique de sur-échantillonnage SVM SMOTE (Tang *et al.*, 2008). À l'inférence, nous appliquons un seuil supplémentaire sur la probabilité renvoyée par les classifieurs (un seuil par classifieur), ce qui nous fournit un contrôle supplémentaire sur le compromis entre précision et rappel. Les seuils que nous utilisons ont été réglés par échantillonnage aléatoire pour maximiser la précision tout en maintenant le nombre d'exemples rejetés inférieur à 70%.

4 Méthode d'évaluation

Au vu de la petite taille des sous-ensembles utilisables pour l'entraînement, nous avons choisi de mettre en œuvre une validation croisée, en suivant un schéma *leave-one-speaker-out* pour mesurer l'adaptabilité des systèmes à de nouveaux locuteurs.

Comme les modèles de reconnaissance de phones utilisés dans l'approche transcription sont limités au jeu de phones qui a servi à leur pré-entraînement, nous les avons entraînés sur des sous-ensembles issus de notre corpus APPR : celui des énoncés ne comportant que des phones du français (APPR-fr) et celui des énoncés ne comportant que des phones du français ou du japonais (APPR-frjp). La table 2 donne des informations sur ces ensembles. Le nombre d'énoncés de APPR-frjp paraît petit par rapport aux 7100 énoncés du corpus, mais c'est parce que les 3600 énoncés contenant des réalisations ambiguës ne sont pas utilisables pour entraîner la reconnaissance de phones.

Sous-ensemble	Durée (h)	#énoncés	#phones
APPR-fr	1,22	2500	17000
APPR-frjp	1,51	3200	21000

TABLE 2 – Taille des données d'entraînement pour l'approche transcription

Les classifieurs utilisés dans l'approche alignement sont entraînés sur les réalisations de /y/ ou de /ʒ/ selon le classifieur. Les seuils des modèles hiérarchiques sont réglés après la validation croisée, à partir des probabilités prédites par les classifieurs pendant la phase de validation.

Nous analysons les résultats de nos systèmes en terme de performances sur la tâche de prédiction de la catégorie didactique. Comme nous sommes limités à deux phonèmes d'intérêt, /y/ et /ʒ/, nous pouvons regarder les performances classe par classe, que nous présentons en terme de précision et de rappel. Pour les systèmes d'EPAO, il est plus dommageable de marquer comme erronée une prononciation correcte que de manquer une prononciation incorrecte (Witt, 2012). Par rapport à nos métriques, cela veut dire que la précision sur les erreurs a plus d'importance que le rappel. Nous nous fixons un objectif de 85 % de précision pour dire si un système a des performances suffisantes pour une catégorie donnée. Par ailleurs, nous considérons qu'avoir des performances insuffisantes sur certaines catégories ayant peu de représentants n'est pas forcément dommageable. Nous ne nous intéressons plus aux catégories « autre » car elles n'ont pas d'intérêt didactique.

5 Résultats

Après avoir évalué les deux modèles de reconnaissance de phones au sein du système basé transcription, et le système basé alignement, nous avons obtenu les résultats consignés dans la table 3. Ils nous informent tout d'abord que l'approche alignement atteint des scores de précision en général plus élevés que l'approche transcription (par exemple 92,7% contre 83,3% sur « /ʒ/ correct »), mais des rappels plus faibles (23,5 % sur cette même classe).

Métrique →	Précision (%)			Rappel (%)		
Approche →	Transcription		Alignement	Transcription		Alignement
Modèle →	Al-fr	Al-frjp		Al-fr	Al-frjp	
/y/ correct	88,5	87,9	**95,0**	**74,1**	73,5	32,5
/y/ perçu comme [ø;œ]	27,3	30,2	**61,3**	39,7	**49,0**	12,6
/y/ perçu comme [j+V]	59,6	52,4	**84,6**	**32,2**	25,3	12,8
/ʒ/ correct	61,1	83,3	**92,7**	**81,5**	75,7	23,5
/ʒ/ perçu comme [dʒ]	83,3	84,4	**88,2**	60,6	**87,6**	31,2
/ʒ/ perçu comme [ʃ]	**25,0**	21,6	0,0	**47,4**	42,1	0,0

TABLE 3 – Résultats des deux approches sur les différentes catégories didactiques

Pour l'approche transcription, les différences entre les deux modèles sont notables. Sur le phonème /ʒ/, l'apprentissage hybride fait diminuer le nombre de confusions de [dʒ] pour [ʒ], ce qui se traduit par des gains absolus de 22,2% de précision sur « /ʒ/ correct » et de 27,0% de rappel sur « /ʒ/ perçu comme [dʒ] ». Sur le phonème /y/, « /y/ perçu comme [ø;œ] » gagne en précision et en rappel (+2,9% et +9,3%) tandis que « /y/ perçu comme [j+Voyelle] » perd en performances (-7,2% et -6,9%), de même que « /y/ correct » dans une moindre mesure, du fait d'une tendance à plus prédire « /y/ perçu comme [ø;œ] ».

L'approche alignement a, sur une majorité de classes, une précision plus élevée et un rappel plus faible que l'approche transcription. Pour les classes où c'est le cas, l'écart absolu de précision va de +3,8% sur « /ʒ/ perçu comme [dʒ] » à +31,1% sur « /y/ perçu comme [ø;œ] », tandis que l'écart de rappel va de -19,4% sur « /y/ perçu comme [j+Voyelle] » à -58,0% sur « /ʒ/ correct ».

Le nombre de représentants a un certain impact sur les résultats : toutes les approches testées ont de meilleurs résultats sur les catégories majoritaires (/y/ et /ʒ/ bien prononcés, « /ʒ/ perçu comme [dʒ] ») que sur les catégories avec moins d'exemples (« /y/ perçu comme [ø;œ] » par exemple).

6 Discussion

Notre objectif de précision de 85 % est atteint pour le phonème /ʒ/ avec le système basé sur l'alignement : 92,7 % et 88,2 % respectivement sur « correct » et « perçu comme [dʒ] ». Le système basé sur la transcription avec Al-frjp est tout juste en dessous de l'objectif (83,3 % et 84,4 %). Prédire avec suffisamment de précision le diagnostic « perçu comme [ʃ] », qui ne représente de toute façon que 1,2 % des réalisations, semble hors de portée des approches évaluées. Pour le phonème /y/, dépasser l'objectif ailleurs que sur la classe « correct », majoritaire, semble difficile. Le système basé sur l'alignement l'atteint presque sur « perçu comme [j+V] » avec 84,6 % de précision, mais de manière surprenante les résultats sont moins bons sur « perçu comme [ø;œ] » (61,3 % de précision), alors que ce diagnostic a 74 % de représentants en plus que « perçu comme [j+V] ».

Notre choix de régler les seuils de nos classifieurs afin de maximiser la précision se voit bien dans les résultats de l'approche alignement. En effet, nous rejetons beaucoup d'exemples (d'où un faible rappel). Si nous déployons ce système dans un programme d'EPAO, un apprenant devra donc réaliser plusieurs exercices avant d'obtenir un diagnostic juste, voire obtenir un diagnostic tout court. C'est particulièrement gênant pour /y/ où le rappel sur les erreurs avoisine $^1/_8$.

Nos résultats montrent que l'apprentissage hybride permet au modèle de reconnaissance de phones de s'améliorer sur la tâche de diagnostic d'erreurs de prononciation. La baisse des confusions entre [ʒ] et [dʒ] s'explique par le fait que le corpus de japonais contient autant d'exemples de ces deux sons. De même, l'intégration du phone [ɯ] du japonais, qui représente une bonne partie des réalisations de /y/ perçues comme [ø;œ], permet de mieux détecter cette catégorie. Dans les deux cas, cette amélioration s'accompagne cependant d'une augmentation des confusions entre les autres classes et la classe qui s'améliore, amoindrissant le gain de précision et diminuant le rappel des autres classes. Si le diagnostic d'erreur s'améliore nettement sur /ʒ/, l'amélioration est moins sensible sur /y/.

7 Conclusions

Cet article compare deux approches de détection et diagnostic d'erreurs de prononciation, l'une basée sur une transcription phonétique et l'autre sur un alignement de signaux audio. Nous les évaluons sur un corpus d'apprenants japonais du français dont les annotations sont guidées par des connaissances didactiques, ce qui rend les diagnostics pertinents pour l'apprentissage des langues. Nous trouvons que l'approche basée sur l'alignement est plus adaptée pour l'EPAO, avec des précisions plus élevées que l'approche basée sur la transcription, imputables au contrôle qu'elle fournit sur le compromis précision/rappel.

Les travaux présentés ici se focalisent sur deux phonèmes particuliers pour une paire L1/L2 donnée, mais la méthodologie mise en œuvre pourrait aisément être appliquée à d'autres phonèmes ou paires de langues, pourvu qu'un corpus de parole non-native intégrant des connaissances didactiques soit disponible. Aussi, il serait intéressant de mener plus d'expérimentations dans ce sens.

Remerciements

Ces travaux ont été financés par l'ANR dans le cadre du LabCom ALAIA (ANR-18-LVC3-001).

Références

ARDILA R., BRANSON M., DAVIS K., HENRETTY M., KOHLER M., MEYER J., MORAIS R., SAUNDERS L., TYERS F. M. & WEBER G. (2020). Common voice : A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, p. 4211–4215.

BAEVSKI A., ZHOU Y., MOHAMED A. & AULI M. (2020). wav2vec 2.0 : A framework for self-supervised learning of speech representations. In H. LAROCHELLE, M. RANZATO, R. HADSELL, M. BALCAN & H. LIN, Éds., *Advances in Neural Information Processing Systems*, volume 33, p. 12449–12460 : Curran Associates, Inc.

BREIMAN L. (2001). Random forests. *Machine learning*, **45**, 5–32.

DETEY S. & RACINE I. (2016). L'apprentissage de la prononciation d'une langue étrangère : le cas du français. In S. DETEY, J. EYCHENNE, Y. KAWAGUCHI & I. RACINE, Éds., *La prononciation du français dans le monde : du natif à l'apprenant*, chapitre 14, p. 84–96. Paris : CLE International.

GELIN L., DANIEL M., PINQUIER J. & PELLEGRINI T. (2021). End-to-end acoustic modelling for phone recognition of young readers. *Speech Communication*, **134**, 71–84.

GHIO A., LALAIN M., GIUSTI L., POUCHOULIN G., ROBERT D., REBOURG M., FREDOUILLE C., LAARIDH I. & WOISARD V. (2018). Une mesure d'intelligibilité par décodage acoustico-phonétique de pseudo-mots dans le cas de parole atypique. In *XXXIIe Journées d'Etudes sur la Parole*, p. 285–293 : ISCA.

GHOSH S., FAUTH C., LAPRIE Y. & SINI A. (2017). End-to-End Acoustic Feedback in Language Learning for Correcting Devoiced French Final-Fricatives. In *Proc. Interspeech 2017*, p. 349–353. DOI : 10.21437/Interspeech.2017-1031.

HARRISON A. M., LO W.-K., QIAN X.-J. & MENG H. (2009). Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training. In *Proc. Speech and Language Technology in Education (SLaTE 2009)*, p. 45–48.

KAMIYAMA T., DETEY S. & KAWAGUCHI Y. (2016). Les japonophones. In S. DETEY, J. EYCHENNE, Y. KAWAGUCHI & I. RACINE, Éds., *La prononciation du français dans le monde : du natif à l'apprenant*, chapitre 24, p. 155–161. Paris : CLE International.

KORZEKWA D., LORENZO-TRUEBA J., DRUGMAN T. & KOSTEK B. (2022). Computer-assisted pronunciation training—speech synthesis is almost all you need. *Speech Communication*, **142**, 22–33. DOI : https ://doi.org/10.1016/j.specom.2022.06.003.

LABORDE V., PELLEGRINI T., FONTAN L., MAUCLAIR J., SAHRAOUI H. & FARINAS J. (2016). Pronunciation Assessment of Japanese Learners of French with GOP Scores and Phonetic Information. In *Proc. Interspeech 2016*, p. 2686–2690. DOI : 10.21437/Interspeech.2016-513.

LEUNG W.-K., LIU X. & MENG H. (2019). Cnn-rnn-ctc based end-to-end mispronunciation detection and diagnosis. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 8132–8136. DOI : 10.1109/ICASSP.2019.8682654.

LI X., DALMIA S., LI J., LEE M., LITTELL P., YAO J., ANASTASOPOULOS A., MORTENSEN D. R., NEUBIG G., BLACK A. W. *et al.* (2020). Universal phone recognition with a multilingual allophone system. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 8249–8253 : IEEE.

LIN B. & WANG L. (2022). Phoneme mispronunciation detection by jointly learning to align. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 6822–6826. DOI : 10.1109/ICASSP43922.2022.9746727.

NEEDLEMAN S. B. & WUNSCH C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**(3), 443–453. DOI : https ://doi.org/10.1016/0022-2836(70)90057-4.

SALVADOR S. & CHAN P. (2007). Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, **11**(5), 561–580.

SANCINETTI M., VIDAL J., BONOMI C. & FERRER L. (2022). A transfer learning approach for pronunciation scoring. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 6812–6816. DOI : 10.1109/ICASSP43922.2022.9747727.

TANG Y., ZHANG Y.-Q., CHAWLA N. V. & KRASSER S. (2008). Svms modeling for highly imbalanced classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, **39**(1), 281–288.

WITT S. M. (2012). Automatic error detection in pronunciation training : Where we are and where we need to go. In *International Symposium on automatic detection on errors in pronunciation training*, p. 1–8.

WITT S. M. & YOUNG S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech communication*, **30**(2-3), 95–108.

WU M., LI K., LEUNG W.-K. & MENG H. (2021). Transformer Based End-to-End Mispronunciation Detection and Diagnosis. In *Proc. Interspeech 2021*, p. 3954–3958. DOI : 10.21437/Interspeech.2021-1467.

XU X., KANG Y., CAO S., LIN B. & MA L. (2021). Explore wav2vec 2.0 for Mispronunciation Detection. In *Proc. Interspeech 2021*, p. 4428–4432. DOI : 10.21437/Interspeech.2021-777.