Détection et classification automatiques d'erreurs de prononciation en L2 : approche basée sur les connaissances didactiques

Romain Contrain, Julien Pinquier, Lionel Fontan, Isabelle Ferrané

Détection et le classification automatiques d'erreurs de prononciation en L2 : approche basée sur les connaissances didactiques.

Romain Contrain[1], Julien Pinquier[1], Lionel Fontan[2], and Isabelle Ferrané[1]

[1]IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France
[2]Archean Labs, Montauban, France

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

Les outils d'Entraînement à la Prononciation Assisté par Ordinateur (EPAO) sont intéressants pour l'apprentissage des langues dans la mesure où la majorité des apprenants n'a pas accès à un professeur particulier pour travailler ces aspects. Ces outils doivent être capables d'effectuer la détection et le diagnostic des 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. La méthode la plus citée est le Goodness of Pronunciation (GOP) [1], mais des méthodes plus récentes emploient des classifieurs basés sur des réseaux de neurones profonds ou des représentations issues de réseaux de neurones comme wav2vec 2.0 [2]. 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. Dans [3] les auteurs effectuent la phase de reconnaissance de phones selon deux architectures basées sur des *Transformers*. Le meilleur système obtient 81,3% de précision et 80,7% de rappel pour la détection et 10,0% d'erreur de diagnostic sur le corpus CU-CHLOE d'apprenants chinois de l'anglais.

Dans le cadre de nos travaux, nous nous plaçons dans le contexte d'une tâche de répétition de mots ou de phrases simples, sur la base de stimuli audio présentés aux apprenants. La prononciation correcte est connue et contient une difficulté à leur faire travailler. Les réalisations probables du phonème cible sont classées dans des catégories didactiques correspondant aux retours qu'un enseignant fournirait selon la présence et le type d'erreurs. Nous explorons deux approches qui se basent sur les deux tendances observées dans la littérature : l'une repose sur un alignement entre la prononciation cible et la production, et l'autre se base sur une transcription phonétique de la production.

La première approche réalise d'abord un alignement entre signaux avec l'algorithme Dynamic Time Warping (DTW), en utilisant des MFCC. Le segment de la production correspondant au phonème cible est ensuite classé dans l'une des catégories didactiques identifiées, en utilisant diverses mesures tirées de la littérature et un classifieur hiérarchique binaire basé sur la méthode des *random forest*. L'intérêt de cette approche est qu'elle est indépendante de la langue.

La seconde approche réalise une transcription phonétique de la production à l'aide d'un réseau récurrent bidirectionnel à mémoire courte et long terme (BiLSTM) [4] pré-entraîné sur de la parole native de la L1 et de la L2 puis adapté à la parole d'apprenants. La transcription est ensuite alignée avec la prononciation cible via un alignement de Needleman-Wunsch [5]. La réalisation correspondant au phonème cible est alors classée dans la catégorie didactique avec laquelle elle est la plus similaire (selon des critères de similarité entre phones inspirés de [6]). Avec cette approche, l'utilisation de *transfer learning* permet de gérer le manque de données d'entraînement.

Notre étude a été réalisée sur un corpus de 7112 énoncés produits par 67 apprenants japonais du français et annotés au niveau phonétique par deux experts. Notre étude se limite pour l'instant aux phonèmes /ʒ/ et /y/, jugés parmi les plus importants pour l'apprentissage du français par des japonophones. Nous disposons ainsi de 1540 réalisations de /ʒ/ et de 1183 réalisations de /y/.

Pour le phonème /ʒ/, les résultats sont encourageants, avec des précisions assez élevées sur les deux catégories les plus fréquentes. Le système basé sur la transcription par un BiLSTM donne

les meilleurs résultats. Nous obtenons ainsi une précision moyenne de 83,8% sur ces catégories. Pour le phonème /y/, les résultats sont moins bons, avec des précisions moins élevées sur les catégories d'erreur les plus fréquentes. Le système basé sur un alignement entre signaux et un classifieur hiérarchique binaire donne les meilleurs résultats. On détecte certes les prononciations correctes avec une précision de 96,7% mais la précision moyenne sur les catégories d'erreurs les plus fréquentes est de seulement 57,7%. Ces différences peuvent s'expliquer en partie par les différences de distribution des catégories entre les phonèmes : pour /ʒ/ le corpus compte 647 prononciations correctes et 862 représentants de la principale catégorie d'erreur, ce qui est assez équilibré. Par contre, pour /y/, il y a 869 prononciations correctes, mais seulement 151 et 87 représentants pour les deux catégories d'erreurs les plus fréquentes.

Cette première étude, son application à une paire de langues donnée (Japonais/Français) et aux phonèmes cibles faisant l'objet de difficultés typiques de cet apprentissage, a permis d'explorer deux approches potentiellement complémentaires pour la détection et de diagnostic d'erreurs de prononciation. Les résultats prometteurs vont se poursuivre par le prise en compte d'autres phonèmes cible. La méthodologie proposée peut également être généralisée à d'autres L1 ou d'autres paires de langues.

Références

[1] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[2] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0 : A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460. [Online]. Available : https ://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

[3] M. Wu, K. Li, W.-K. Leung, and H. Meng, "Transformer Based End-to-End Mispronunciation Detection and Diagnosis," in *Proc. Interspeech 2021*, 2021, pp. 3954–3958.

[4] X. Li, S. Dalmia, J. Li, M. Lee, P. Littell, J. Yao, A. Anastasopoulos, D. R. Mortensen, G. Neubig, A. W. Black *et al.*, "Universal phone recognition with a multilingual allophone system," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8249–8253.

[5] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970. [Online]. Available : https ://www.sciencedirect.com/science/article/pii/0022283670900574

[6] A. Ghio, M. Lalain, L. Giusti, G. Pouchoulin, D. Robert, M. Rebourg, C. Fredouille, I. Laaridh, and V. Woisard, "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*. ISCA, 2018, pp. 285–293.