Réseau de neurones convolutif pour l'évaluation automatique de la prononciation

Thomas Pellegrini[1] Lionel fontan[2] Halima Sahraoui[3]
(1) IRIT - Université de Toulouse, 31062, Toulouse, France
(2) Archean Technologies, 1899 av. d'Italie, 82000, Montauban, France
(3) Octogone-Lordat - Université de Toulouse, 31058, Toulouse, France
thomas.pellegrini@irit.fr, lfontan@archean.fr, sahraoui@univ-tlse2.fr

RÉSUMÉ

Dans cet article, nous comparons deux approches d'évaluation automatique de la prononciation de locuteurs japonophones apprenant le français. La première, l'algorithme standard appelé *Goodness Of Pronunciation* (GOP), compare les vraisemblances obtenues lors d'un alignement forcé et lors d'une reconnaissance de phones sans contrainte. La deuxième, nécessitant également un alignement préalable, fait appel à un réseau de neurones convolutif (CNN) comme classifieur binaire, avec comme entrée des trames de coefficients spectraux. Les deux approches sont évaluées sur deux phonèmes cibles /R/ et /v/ du français, particulièrement difficiles à prononcer pour des Japonophones. Les paramètres du GOP (seuils) et du CNN sont estimés sur un corpus de parole lue par des locuteurs natifs du français, dans lequel des erreurs de prononciation artificielles sont introduites. Un gain de performance relatif de 13,4% a été obtenu avec le CNN, avec une précision globale de 72,6%, sur un corpus d'évaluation enregistré par 23 locuteurs japonophones.

ABSTRACT

CNN-based automatic pronunciation assessment of Japanese speakers learning French

In this paper, we compare two approaches for the automatic evaluation of the pronunciation of Japanese speakers speakers learning French. The first one, the standard algorithm called Goodness Of Pronunciation (GOP), compares likelihoods obtained during forced alignment and during phone recognition with no constraint. The second, also requiring a signal-to-phone alignment, uses a convolutional neural network (CNN) as a binary classifier, with frames of spectral coefficients as input. Both approaches are evaluated on two target French phonemes /R/ and /v/, particularly difficult to pronounce for Japanese-speaking natives. GOP decision thresholds and CNN parameters are estimated on a read speech corpus of French native speakers, in which artificial pronunciation errors are introduced. A 13.4% performance gain relative was obtained with the CNN, with an overall accuracy of 72.6%, on an evaluation corpus recorded by 23 Japanese native speakers.

MOTS-CLÉS : évaluation automatique de la prononciation, réseau de neurones convolutif, français langue étrangère.

KEYWORDS: Automatic pronunciation assessment, convolutional neural network, CNN, Goodness-of-Pronunciation, French as a second language.

1 Introduction

Les systèmes d'apprentissage d'une langue seconde assisté par ordinateur tentent d'évaluer automatiquement la prononciation pour aider les apprenants. Pour l'évaluation au niveau segmental, une approche standard consiste à attribuer un score de prononciation à chaque réalisation de phone (Eskenazi, 2009). Deux grands types d'approches peuvent être distingués : 1) les approches fondées sur des scores de systèmes de reconnaissance, bruts (Sevenster *et al.*, 1998), ou relatifs, sous forme de rapports de vraisemblances comme c'est le cas de l'algorithme *Goodness Of Pronunciation* (GOP) (Witt, 1999; Kanters *et al.*, 2009) utilisé dans la présente étude, 2) les approches « signal », qui font appel à des classifieurs prenant en entrée des paramètres acoustiques (Strik *et al.*, 2007).

Dans (Strik *et al.*, 2007) par exemple, les auteurs comparent ces deux types d'approches pour distinguer les deux phonèmes /k/ et /x/ sources de confusion pour les apprenants du néerlandais. Ils obtiennent des performances légèrement meilleures en utilisant une analyse discriminante linéaire (LDA) avec des paramètres acoustiques. Cette approche présente néanmoins le désavantage d'être spécifique à chaque phone que l'on souhaite évaluer et il faut ré-estimer les poids de la LDA à chaque nouveau phone cible. Dans l'étude présentée dans cet article, nous comparons également les deux types d'approches, en utilisant un réseau de neurones convolutif (CNN), qui, outre ses performances remarquables en reconnaissance automatique de la parole, permet d'évaluer plusieurs phones simultanément.

Plus précisément, nous comparons les approches suivantes : 1) une variante plus robuste de l'algorithme de base GOP, appelée f-GOP, proposée par Luo *et al.* (2010), avec et sans seuils (utilisation d'une régression logistique), 2) un réseau de neurones convolutif qui prend en entrée des paramètres acoustiques (coefficients F-BANK) ainsi que l'identité du phone attendu, et qui prend une décision binaire trame à trame d'acceptation ou de rejet d'une prononciation. Une décision finale pour un phone est prise par un simple vote majoritaire sur l'ensemble des trames du phone.

Les expériences d'évaluation ont été réalisées sur le corpus PHON-IM, corpus de parole produite dans une tâche de répétitions de mots, recueillie auprès de 23 locuteurs japonais qui apprennent le français comme langue étrangère (FLE). Nous avons centré cette étude sur deux phones particulièrement difficiles à maîtriser pour les japonophones : [R] et [v], souvent confondus avec [l] et [b] respectivement (Tomimoto & Takaoka, 2008; Yamasaki & Hallé, 1999). La taille de ce corpus étant très petite pour servir à la fois à l'apprentissage et à l'évaluation des méthodes, nous avons utilisé le corpus de parole lue BREF80, enregistré par 80 locuteurs français natifs. Des erreurs de prononciation sont simulées en introduisant des substitutions de phones dans le dictionnaire de prononciation utilisé pour réaliser les alignements. Si cette méthode a été appliquée avec succès dans Kanters *et al.* (2009), ce travail nous a permis d'en déceler des limites que nous décrirons.

2 Approches

Les approches utilisées nécessitent d'aligner le signal de parole avec les séquences de phones attendus. Nous avons utilisé pour cela des modèles acoustiques de phones indépendants du contexte (39 monophones), plus adaptés que des modèles dépendants du contexte pour des applications de détection d'erreurs de prononciation (Kawahara & Minematsu, 2012). Ces modèles sont des HMM gauche-droite à trois états avec des mélanges de Gaussiennes à 32 composantes, entraînés sur le corpus ESTER phase I (de Calmès *et al.*, 2005). Ce sous-corpus contient 31 heures de parole de diverses émissions de radio française. Les modèles sont disponibles en ligne (Farinas, 2013).

2.1 Algorithme *f-GOP*

Cette méthode a été proposée au départ pour l'évaluation de prononciation non-native (Witt & Young, 2000; Kanters *et al.*, 2009; Luo *et al.*, 2010), et a été également utilisée avec succès pour caractériser les troubles pathologiques de production de la parole dans des cas de sévérité modérée (Pellegrini *et al.*, 2015).

L'algorithme GOP de référence peut être décomposé en trois étapes : 1) la phase d'alignement forcé d'une séquence de phones attendus (parole lue) au signal de parole, 2) la phase de reconnaissance de phones sans contrainte et 3) le calcul des scores comme la différence entre les log-vraisemblances des deux phases précédentes pour chaque phone aligné. Les scores varient entre 0 et 10 environ, et plus la valeur est grande, plus une erreur de prononciation est susceptible d'avoir été déctectée. Les ordres de grandeur des vraisemblances dépendent entre autres du phone considéré. Pour cette raison, il est commun de déterminer à partir d'un sous-corpus de développement les seuils de décision pour chaque phone cible.

Dans ce travail, nous utilisons une version plus performante du GOP, appelée f-GOP, qui force la phase de reconnaissance libre à utiliser les segments trouvés lors de l'alignement (Luo *et al.*, 2010). Les seuils d'acceptabilité des phones [R] et [v] ont été estimés à 1,13 et 2,97 respectivement. Pour déterminer ces seuils, nous avons utilisé BREF80, le corpus de parole lue par des locuteurs français natifs, que nous décrivons à la section 3. Les locuteurs de ce corpus étant des francophones natifs, nous considérons toutes les occurrences des deux phones comme acceptables (classe positive). Pour simuler des erreurs de prononciation, nous considérons que toutes les occurrences de [l] et [b] correspondent à des prononciations erronées de /R/ et /v/ respectivement. Pour ce faire, nous forçons le système d'alignement à utiliser un [R] à la place des [l] en remplaçant le phone [l] par [R] dans le lexique de prononciation (même chose pour [b] et [v]).

2.2 Régression logistique (f-GOP+RL)

Pour éviter de devoir fixer des seuils de décision, nous avons utilisé un classifieur fondé sur une régression logistique (RL). Très populaire en traitement du langage naturel, cette technique obtient des performances similaires aux séparateurs à vaste marge (Theodoridis, 2015), avec l'avantage de mettre en jeu des poids θ qui ont une interprétation sur l'importance des paramètres d'entrée. Pour pouvoir comparer les performances du modèle RL avec f-GOP, nous avons utilisé comme paramètres d'entrée uniquement les scores f-GOP et l'identité du phone attendu. Les poids du modèle sont entraînés sur les mêmes exemples de BREF80 qui ont servi à fixer les seuils des scores f-GOP. On constate que le poids du score GOP après apprentissage vaut -0,633, une valeur négative qui correspond bien au fait que plus un score GOP est élevé, plus une erreur de prononciation est vraisemblable. Les poids attribués aux paramètres catégoriels d'identité du phone sont 0,627 et 0,445 pour /v/ et /R/ respectivement. Le poids pour /v/ est légèrement plus grand que celui de /R/, ce qui est également cohérent avec le fait que le seuil GOP trouvé précédemment est plus élevé pour ce phone.

2.3 Réseau de neurones convolutif (CNN)

La figure 1 illustre l'architecture du réseau mis en place pour cette étude. Il comporte une couche d'entrée composée d'une trame de 40 coefficients log-F-BANK calculés sur une fenêtre de 20ms, à

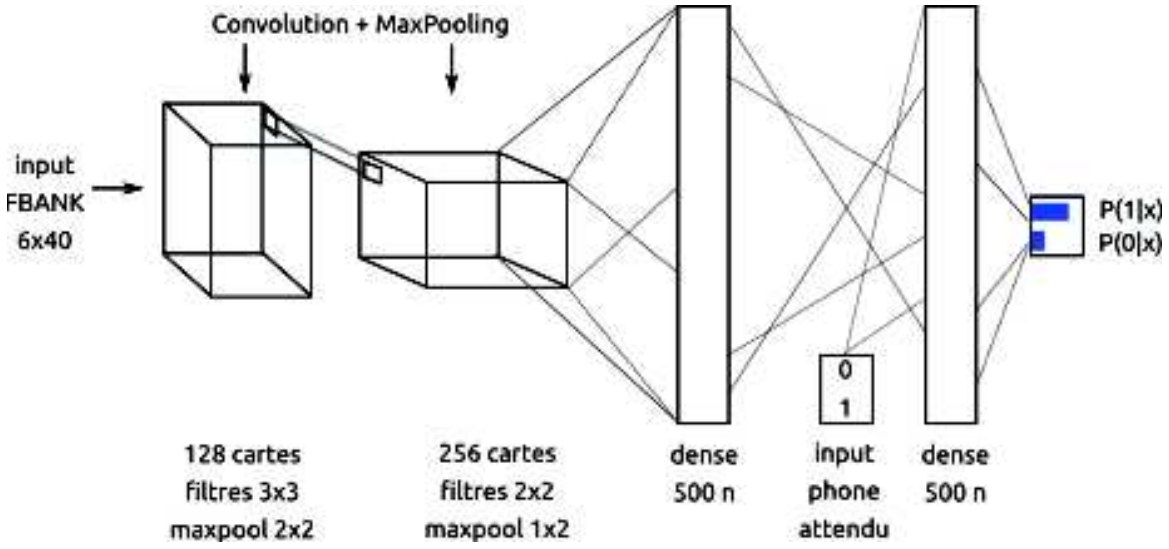

FIGURE 1 – Architecture du réseau convolutif.

laquelle ont été ajoutées les trois trames voisines précédentes et les deux trames voisines suivantes pour un total de six trames. Deux trames voisines sont séparées par 10ms. Deux couches de convolution avec sous-échantillonnage (*Max pooling*) permettent d'obtenir respectivement 128 et 256 cartes d'activation qui servent de paramètres d'entrée à deux couches cachées denses de 500 neurones avec une fonction d'activation ReLu. L'information de l'identité du phone attendu, connue grâce à un alignement forcé préalable (GOP, phase 1), est donnée à la dernière couche cachée de décision, sous la forme d'un vecteur de zéros avec un unique 1 (représentation *one-hot*). Cette information est nécessaire puisqu'un seul modèle est utilisé pour tous les phones d'intérêt (ici au nombre de deux : [R] et [v]). Une décision par trame est prise et la décision globale pour un phone est prise à l'aide d'un vote majoritaire sur l'ensemble des trames qui le compose. Les poids du modèle ont été initialisés à l'aide de la méthode « Xavier » (Glorot & Bengio, 2010), et entraînés avec une descente de gradient momentum Nesterov, avec une fonction de coût de type entropie binaire croisée. La méthode de régularisation *dropout* ($p = 0,5$) n'est utilisée qu'avec les couches cachées denses. D'autres architectures ont été testées, avec une seule couche de convolution ou avec trois couches de convolution par exemple, et celle présentée ici a donné les meilleurs résultats. Pour la mise en œuvre de ces modèles, nous avons utilisé les boîtes à outils Theano (Bergstra *et al.*, 2010) et Lasagne [1].

Pour entraîner le modèle, nous avons divisé le corpus BREF80, décrit dans la section suivante, en deux sous-corpus *Train* et *Val* dans les proportions 90% / 10%, soit 300K / 30K exemples, respectivement. La figure 2 montre l'évolution du coût sur *Train* et *Val*, ainsi que la précision obtenue sur *Val*, au cours des 100 premières itérations d'apprentissage. Une droite horizontale situe la performance obtenue sur PHON-IM par le modèle final : 88,9% de classification correcte des trames. Si le coût sur *Train* continue à diminuer après 100 itérations, la performance sur *Val* atteint un plateau rapidement. Le critère d'arrêt que nous avons utilisé était une diminution minimale de 1e-3 sur le coût calculé sur *Val* au cours de trois itérations successives, ce qui a conduit à un nombre de 178 itérations.

1. https://github.com/Lasagne/Lasagne

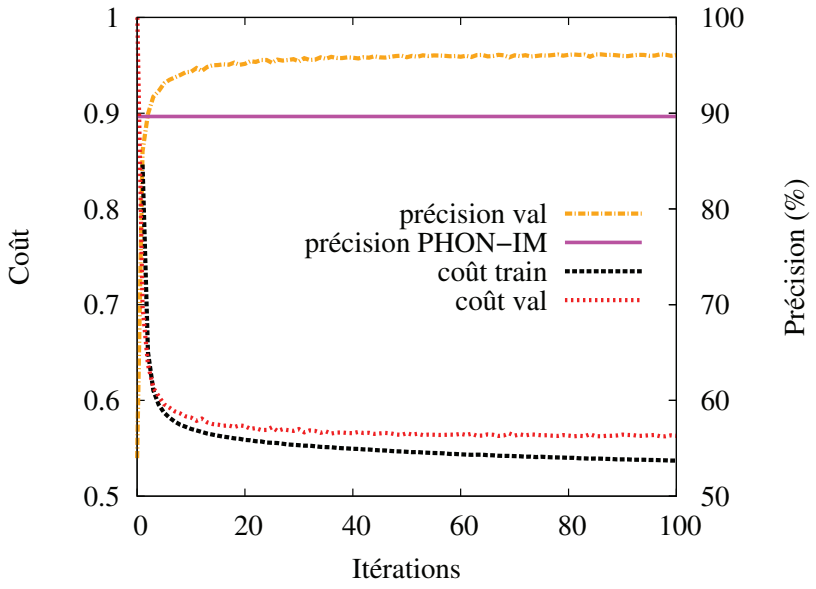

FIGURE 2 – Évolution du coût sur *train*, du coût et précision sur *val*, au cours des 100 premières itérations d'apprentissage. La performance sur le corpus de test PHON-IM est indiquée par une droite horizontale (88,9%).

corpus	BREF80		PHON-IM	
	correct	incorrect	correct	incorrect
/R/	21K	16K (phone [l])	215	128
/v/	5K	3K (phone [b])	267	50

TABLE 1 – Nombres d'occurrences de /R/ et de /v/ dans BREF80 et PHON-IM.

3 Corpora

3.1 BREF80

Le corpus BREF est un corpus de plus de 100 heures de parole lue, recueillie auprès de 120 locuteurs français natifs, qui ont lu des textes du journal *Le Monde* (Gauvain *et al.*, 1990). Nous en avons utilisé une sous-partie, appelée BREF80, qui correspond aux enregistrements de 80 locuteurs.

Le nombre d'occurrences des deux phones cibles [R] et [v] dans ce corpus est indiqué dans le tableau 1 : 21K et 5K. Toutes ces occurrences sont considérées comme des réalisations correctes. Les nombres d'occurrences incorrectes correspondent aux nombres de [l] et [b] du corpus, 16K et 3K, considérés comme des exemples de la classe négative. Comme dit dans la section précédente, ce corpus a été réparti en deux sous-corpus *Train* et *Val* dans les proportions 90% / 10%, spécifiquement pour l'apprentissage du modèle CNN. Cette division est nécessaire pour établir un critère d'arrêt sur les itérations de descente de gradient.

TABLE 2 – Corpus PHON-IM : jugements d'acceptabilité par phonème et par position

	/v/		/R/	
	Acceptable	Non accept.	Acceptable	Non accept.
Initiale	89	30 (25,2%)	53	59 (52,7%)
Intervocalique	75	6 (7,4%)	56	55 (49,6%)
Finale	103	14 (12,0%)	106	14 (11,7%)
Total	277	50 (15,8%)	215	128 (37,3%)

3.2 PHON-IM

3.2.1 Enregistrements utilisés dans le cadre de cette étude

PHON-IM est un projet de recherche visant à étudier les compétences de perception et de production phonétique chez des apprenants japonophones de FLE, et ce dans une perspective longitudinale. Il s'inscrit dans le cadre d'un programme annuel d'échange d'étudiants entre l'Université Ritsumeikan de Kyoto et le Département de FLE (DEFLE) de l'Université Toulouse II – Jean Jaurès. Chaque année un groupe d'apprenants débutants (A1/A2) vient à Toulouse pour un séjour d'immersion linguistique d'un mois. Les réalisations cibles /R/, /l/, /v/ et /b/ présentent typiquement des difficultés particulières pour les locuteurs natifs japonophones en écoute et prononciation du français langue étrangère.

Pour ce travail nous avons utilisé un sous-ensemble d'enregistrements collectés auprès de 23 apprenants japonais, dont la tâche était de répéter des mots ou pseudo-mots dissyllabiques contenant les phonèmes /R/ et /v/. Les deux phonèmes apparaissent à la fois dans les positions initiale, intervocalique et finale, la fréquence d'occurrence dans ces trois positions étant équilibrée au sein du corpus. Au total, le sous-ensemble comprend 368 réalisations de /v/ et 414 réalisations de /R/.

3.2.2 Annotations par des enseignants de FLE

Deux enseignants de FLE ont évalué les 782 réalisations, en indiquant si selon eux la réalisation était acceptable en fonction de la cible attendue. L'accord entre les deux annotateurs est de 84,4% ; cet accord est légèrement supérieur pour les réalisations du phonème /R/ (86,2%) que pour celles du phonème /v/ (82,9%).

En cas de réalisation non acceptable, les deux annotateurs étaient enjoints à indiquer quel était le phone se rapprochant le plus de la réalisation de l'apprenant. Pour le phonème /R/, les réalisations jugées non acceptables ont le plus souvent été décrites comme proches du phone [h] présent dans le système phonético-phonologique du japonais. Pour les réalisations du phonème /v/, c'est la fricative bilabiale [β] qui a le plus souvent été évoquée par les deux annotateurs, et dans une moindre mesure l'occlusive [b].

Au total, parmi les 660 occurrences pour lesquelles un accord inter-annotateur a été obtenu, 15,8% des réalisations de /v/ ont été jugées non acceptables, contre 37,3% pour le /R/. La table 2 décrit les

		Correctement Acceptés			Correctement Rejetés		
Modèles	tx global	prec.	rappel	F1	prec.	rappel	F1
f-GOP	68,5/58,7	73,2/91,5	78,6/56,2	75,8/69,6	58,9/23,5	51,6/72,0	55,0/35,4
f-GOP+RL	71,1/57,1	71,6/92,3	89,3/53,6	79,5/67,8	69,3/23,5	40,6/76,0	51,2/35,9
CNN	68,5/77,0	71,1/91,1	83,7/80,5	76,9/85,5	61,1/35,8	43,0/58,0	50,5/44,3

TABLE 3 – Résultats obtenus sur le corpus de test PHON-IM. Dans chaque cellule, les pourcentages sont donnés pour les phonèmes /R/ et /v/, respectivement.

scores par position. Il semble que ce soit la position initiale qui pose le plus de difficultés, de manière beaucoup plus marquée pour /v/ que pour /R/, avec 30 occurrences jugées non-acceptables pour cette position, contre 6 et 14 en position intervocalique et finale respectivement.

4 Résultats

Dans le tableau 3 donne les mesures de performance obtenues sur le corpus d'évaluation PHON-IM, avec le détail pour les deux phonèmes /R/ et /v/ dans chaque cellule du tableau. De manière globale, les performances f-GOP et f-GOP+RL sont très proches, ce qui est confirmé par le fait que leurs prédictions sont identiques à 89,2% pour /R/ et à 97,2% pour /v/. Ce résultat était attendu dans la mesure où les mêmes informations sont utilisées en entrée, à savoir le score GOP et l'identité du phone attendu. Le modèle de régression linéaire apprend de manière autonome les seuils d'acceptabilité, ce qui confirme l'intérêt d'utiliser un tel classifieur.

En revanche, les prédictions obtenues avec le CNN différent de manière significative : les pourcentages de prédictions identiques avec le modèle f-GOP+LR tombent à 76,4% et 50,5% des cas pour /R/ et /v/, respectivement. La première colonne du tableau donne le taux global de bonne classification, i.e. le ratio du nombre de réalisations correctement acceptées (CA) ou rejetées (CR) sur le nombre d'occurrences total. Les trois approches ont donné un taux similaire pour /R/, autour de 70%. En revanche, le CNN a un taux bien meilleur pour /v/ : 77,0% contre 58,7% et 57,1% pour f-GOP et f-GOP+RL. Cela est essentiellement dû à un rappel des CA meilleur avec CNN, de 80,5%, par rapport aux rappels de f-GOP, 56,2%, et de f-GOP+RL, 53,6%. f-GOP et f-GOP+RL ont tendance à rejeter des occurrences de /v/ qui ont été jugées correctes par les annotateurs. Cette tendance est également visible de par les taux de rappel des occurrences CR, qui sont meilleurs pour ces deux approches, mais qui ne peuvent compenser les valeurs plus faibles des rappels CA et de la précision CR (23,5%).

Pour analyser ces résultats plus en détails, la figure 3 illustre les prédictions faites par le CNN en fonction de la position du phone dans le mot (intiale, intervocalique ou finale) pour /v/ (à gauche) et pour /R/ (à droite). Deux histogrammes de trois barres chacun sont donnés pour chaque position : les occurrences « acceptées » et les occurrences « rejetées ». Les trois barres correspondent de gauche à droite : à l'annotation manuelle, aux nombres d'occurrences correctement (CA et CR) et incorrectement (FA et FR) classées par le CNN. Il est intéressant de noter que pour les deux phones, c'est la position initiale qui est la source du plus grand nombre d'erreurs : les rejets incorrects pour /v/, les acceptations incorrectes pour /R/. La position intervocalique est celle qui a les meilleures performances. Ces différences de qualité des prédictions indiquent que la méthode de simulation d'erreurs devrait prendre en compte la position des phones. Ces résultats semblent montrer que les difficultés de prononciation de /v/ et /R/ des Japonophones ne sont pas les mêmes selon la position du

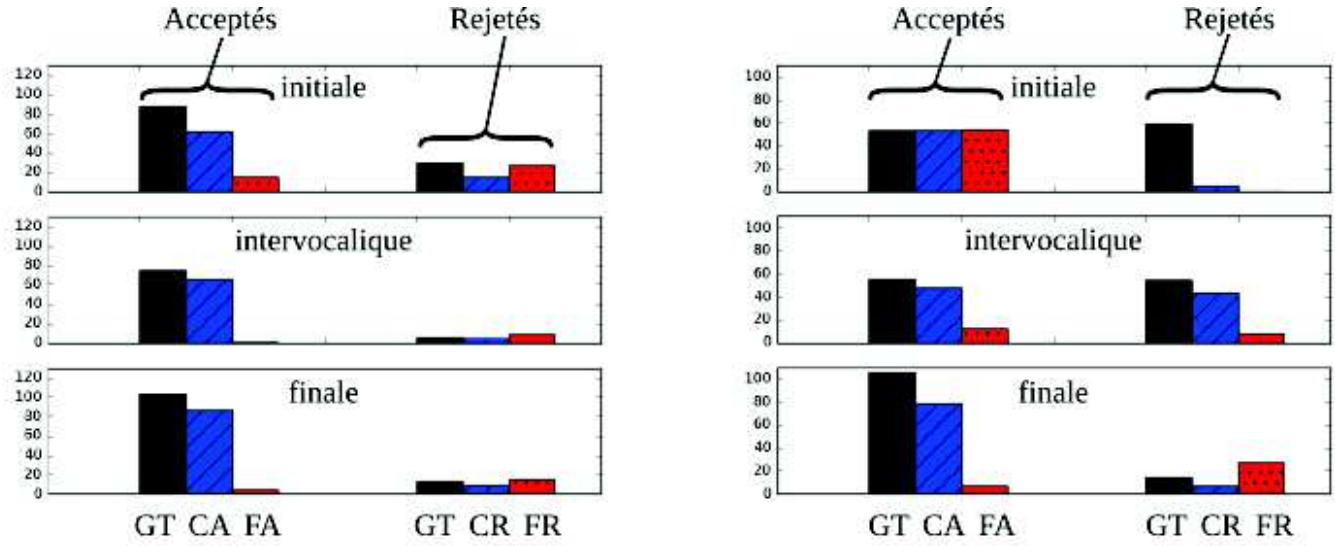

FIGURE 3 – Nombre d'occurrences de /v/ (à gauche) et de /R/ (à droite) acceptées ou rejetées par les annotateurs (noir, GT pour *ground-truth*), correctement acceptés ou rejetés par le CNN (bleu hachuré, CA et CR), et faussement acceptés ou rejetés par le CNN (rouge avec points, FA et FR).

phone, ce qui transparaît dans le fait que les annotateurs ont rejetés moins d'occurrences des phones en position intervocalique que dans les deux autres positions cumulées.

5 Conclusions

Nous avons comparé deux approches d'évaluation automatique de la prononciation de locuteurs japonophones apprenant le français : l'algorithme f-GOP et un réseau de neurones convolutif. Elles ont été évaluées sur deux phones cibles [R] et [v] du français, particulièrement difficiles à prononcer pour des locuteurs japonophones débutants. Un gain de performance relatif de 13,4% a été obtenu avec le CNN, avec une précision globale de 72,6%, sur un corpus de mots recueilli auprès de 23 locuteurs japonophones.

Pour pouvoir mettre en place ces méthodes, nous avons dû recourir à la simulation d'erreurs de prononciation dans un corpus de parole native pour pallier le manque de données de parole d'apprenants. Nous avons constaté que cela présente des limites dans la mesure où des connaissances *a priori* sont nécessaires sur les confusions les plus fréquentes faites par les apprenants. De plus, les différences de performance en fonction de la position intermot des consonnes montrent que simuler les erreurs sans en tenir compte est une approximation qui peut être améliorée.

Nous envisageons de comparer les résultats présentés ici avec deux autres situations : 1) avec une prise en compte plus fine des confusions faites en fonction de la position des phones pour la simulation des erreurs, 2) à l'inverse, sans connaissance *a priori* sur les confusions fréquentes faites par les locuteurs. Dans la deuxième situation, nous envisageons d'utiliser des occurrences de phones du français choisis aléatoirement pour simuler des erreurs, en limitant le nombre d'occurrences pour obtenir un jeu d'apprentissage équilibré. Des améliorations du modèle CNN lui-même sont également envisagées. Enfin, une nouvelle collection de données d'apprenants japonophones est actuellement en cours, dans les mêmes conditions que celles du corpus PHON-IM, ce qui permettra de doubler la taille du corpus d'évaluation.

Références

BERGSTRA J., BREULEUX O., BASTIEN F., LAMBLIN P., PASCANU R., DESJARDINS G., TURIAN J., WARDE-FARLEY D. & BENGIO Y. (2010). Theano : a CPU and GPU math expression compiler. In *Proc. of the Python for Scientific Computing Conference (SciPy)*.

DE CALMÈS M., FARINAS J., FERRANÉ I. & PINQUIER J. (2005). Campagne ESTER : une première version d'un système complet de transcription automatique de la parole grand vocabulaire (Atelier ESTER, Avignon, 30/03/2005-31/03/2005). `http://www.afcp-parole.org/camp_eval_systemes_transcription/private/atelier_mars_2005/pdf/IRIT_ester2005.pdf`. [Online ; accessed 10-April-2016].

ESKENAZI M. (2009). An overview of spoken language technology for education. *Speech Communication*, **51**(10), 832–844.

FARINAS J. (2013). Multilingual phonetic decoders. `http://www.irit.fr/recherches/SAMOVA/pagedap.html`. [Online ; accessed 20-September-2015].

GAUVAIN J.-L., LAMEL L. & ESKENAZI M. (1990). Design considerations and text selection for BREF, a large french read-speech corpus. In *Proc. ICSLP-90*, p. 1097–2000.

GLOROT X. & BENGIO Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS'10)*.

KANTERS S., CUCCHIARINI C. & STRIK H. (2009). The Goodness of Pronunciation Algorithm : A Detailed Performance Study. In *SLaTE 2009 - 2009 ISCA Workshop on Speech and Language Technology in Education*, p. 2–5.

KAWAHARA T. & MINEMATSU N. (2012). *Tutorial on CALL Systems at Interspeech*. Portland.

LUO D., QIAO Y., MINEMATSU N., YAMAUCHI Y. & HIROSE K. (2010). Regularized-MLLR speaker adaptation for computer-assisted language learning system. In *Proc. Interspeech*, p. 594–597, Makuhari.

PELLEGRINI T., FONTAN L., MAUCLAIR J., FARINAS J., ALAZARD-GUIU C., ROBERT M. & GATIGNOL P. (2015). Automatic Assessment of Speech Capability Loss in Disordered Speech. *ACM Trans. Access. Comput.*, **6**(3), 8 :1–8 :14.

SEVENSTER B., KROM G. D. & BLOOTHOOFT G. (1998). Evaluation and training of second-language learners' pronunciation using phoneme-based HMMs. In *Proc. STiLL*, p. 91–94, Marholmen.

STRIK H., TRUONG K. P., DE WET F. & CUCCHIARINI C. (2007). Comparing classifiers for pronunciation error detection. In *Proc. INTERSPEECH*, p. 1837–1840.

THEODORIDIS S. (2015). *Machine Learning*. Elsevier.

TOMIMOTO J. & TAKAOKA Y. (2008). Le français, une langue imprononçcable pour les Japonais ? *Rencontres Pédagogiques du Kansaï*.

WITT S. (1999). *Use of Speech Recognition in Computer-Assisted Language Learning*. Phd dissertation, University of Cambridge, Dept. of Engineering.

WITT S. & YOUNG S. (2000). Phone-level Pronunciation Scoring and Assessment for Interactive Language Learning. *Speech Communication*, **30**, 95–108.

YAMASAKI H. & HALLÉ P. (1999). How do native speakers of japanese discriminate and categorize french /r/ and /l/ ? In *Proceedings of ICPhS*, p. 909–912, San Francisco.