Améliorations d'un système Transformer de reconnaissance de phonèmes appliqué à la parole d'enfants apprenants lecteurs

Lucile Gelin, Thomas Pellegrini, Julien Pinquier, Morgane Daniel

▶ **To cite this version:**

Lucile Gelin, Thomas Pellegrini, Julien Pinquier, Morgane Daniel. Améliorations d'un système Transformer de reconnaissance de phonèmes appliqué à la parole d'enfants apprenants lecteurs. 34èmes Journées d'Études sur la Parole - Parole, Geste, Musique : des unités à leur organisation (JEP 2022), Association Francophone de la Communication Parlée, Jun 2022, Noirmoutier, France. à paraître. hal-03898401

HAL Id: hal-03898401
https://hal.science/hal-03898401

Submitted on 14 Dec 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Améliorations d'un système Transformer de reconnaissance de phonèmes appliqué à la parole d'enfants apprenants lecteurs

Lucile Gelin[1,2] Thomas Pellegrini[1], Julien Pinquier[1], Morgane Daniel[2]
(1) IRIT, Université Paul Sabatier, CNRS, Toulouse, France
(2) Lalilo, 236 rue du Faubourg Saint-Martin, 75010 Paris, France
{lucile.gelin, thomas.pellegrini, julien.pinquier}@irit.fr,
morgane@lalilo.com

RÉSUMÉ

Les performances des systèmes de reconnaissance automatique de la parole d'enfants n'égalent pas celles des systèmes adultes : la parole d'enfant est difficile à reconnaître et peu de données sont disponibles en Français. Nous faisons de plus face ici à la présence d'erreurs de lecture de jeunes enfants. Nous adaptons un système Transformer end-to-end à la parole d'enfants apprenant·e·s lecteur·rice·s. Un entraînement multi-objectif avec une fonction *Connectionist Temporal Classification* (CTC) et un décodage joint CTC/attention réduit le taux d'erreur phonème (PER) de 22,9% à 19,6%. Nous combinons également une augmentation par ajout de bruit de salle de classe et une augmentation innovante par simulation d'erreurs de lecture pour améliorer la robustesse du système, et atteignons un PER de 15,1%. Des analyses détaillées montrent que le système est plus robuste au bruit, et que l'ajout de la fonction CTC et l'augmentation d'erreurs synthétiques aident à mieux reconnaître les erreurs des enfants.

ABSTRACT

Improving a Transformer-based phoneme recognition system for children learning to read

The performance of automatic speech recognition systems for children is not equal to that of systems for adults : child speech is difficult to recognize and few data are available in French. Moreover, we face here the presence of reading errors of young children. We adapt an end-to-end Transformer system to the speech of children learning to read. Multi-objective training with CTC and joint CTC/attention decoding reduce the phoneme error rate (PER) from 22.9% to 19.6%. We also combine a classroom babble noise augmentation and an innovative augmentation consisting in simulating reading mistakes to improve the robustness of our system and achieve a PER of 15.1%. Detailed analyses show that the system is more robust to noise and that the use of CTC and synthetic mistakes augmentation help better recognize children's reading mistakes.

MOTS-CLÉS : parole d'enfant, Transformer, CTC, augmentation de données, bruit de salle de classe, erreurs de lecture synthétiques.

KEYWORDS: child speech, Transformer, CTC, data augmentation, classroom babble noise, synthetic reading mistakes.

1 Introduction

La parole d'enfants de 5 à 7 ans est sujette à des particularités liées à la croissance de leur appareil de production de la parole et au mauvais contrôle de leur corps : mécanismes articulatoires instables et variabilité spectrale intra- et inter-locuteur·rice·s (Lee *et al.*, 1999), fréquences fondamentale et formantiques plus hautes (Mugitani & Hiroya, 2012), erreurs phonologiques (Fringi *et al.*, 2015), etc. Ces différences morphologiques et phonologiques sont les principales causes des faibles performances des systèmes de reconnaissance automatique de la parole (RAP) sur les voix d'enfants.

Les tuteurs numériques de lecture ont un fort impact pédagogique sur les enfants qui apprennent à lire, et plusieurs projets ont vu le jour au fil des ans (Mostow & Aist, 2001; Bolaños *et al.*, 2011; Godde *et al.*, 2017). Lalilo [1] propose un assistant de lecture pour les enfants de 5 à 7 ans, avec un exercice de lecture orale qui fournit un retour personnalisé grâce un système de reconnaissance automatique de phonèmes que nous présentons dans cet article. Travailler sur la parole de lecteur·rice·s non-expert·e·s ajoute des difficultés dues à la présence d'erreurs de lecture : hésitations, faux départs, répétitions et erreurs de déchiffrage (Potamianos & Narayanan, 1998; Yeung & Alwan, 2018).

Des études antérieures sur la RAP des enfants ont démontré que les performances sont inférieures à celles de la parole des adultes (Potamianos & Narayanan, 2003; Shivakumar & Georgiou, 2020; Yeung & Alwan, 2018). Des systèmes hybrides ont montré des améliorations par un entraînement sur un mélange de données d'adulte et d'enfant (Serizel & Giuliani, 2014), ou grâce à un apprentissage par transfert (*Transfer Learning*, TL) (Shivakumar & Georgiou, 2020). Les architectures récentes dites bout-à-bout, ou *end-to-end*, ont atteint les performances des architectures hybrides sur de la parole d'adulte (Karita *et al.*, 2019b) mais sont encore peu courantes pour la RAP d'enfants en raison du peu de données disponibles.

Cet article est une extension de (Gelin *et al.*, 2021b). Les nouveautés concernent l'utilisation d'un décodage joint entre les deux branches CTC et attention du modèle, des expériences d'augmentation des données par l'ajout de bruit de brouhaha (*babble noise*), ainsi que l'amélioration de notre technique d'augmentation par simulation d'erreurs de lecture.

2 Jeux de données

2.1 Parole d'adulte : Common Voice

Le corpus Common Voice [2] est créé via une plateforme participative en ligne, où chacun peut s'enregistrer en train de lire des phrases. En français, le jeu d'entraînement que nous avons utilisé pour ces expériences contient environ 150 heures de parole d'adulte. Chaque enregistrement est validé par deux annotateur·rice·s, le corpus contient donc peu d'erreurs de lecture.

1. https://www.lalilo.com/
2. Corpus disponible sur : https://voice.mozilla.org/fr

2.2 Parole d'enfant : Lalilo

Le corpus Lalilo contient des enregistrements d'enfants du CP au CE2, âgés de 5 à 8 ans, lisant oralement des mots isolés, des phrases et des histoires courtes. Les enregistrements sont principalement recueillis dans le cadre de l'exercice de lecture orale de la plateforme Lalilo, qui est le plus souvent utilisé dans des salles de classe sous surveillance réduite : ils contiennent des niveaux variables de bruit de brouhaha. Les ensembles d'entraînement et de validation contiennent respectivement 13 et 0,41 heures de données. Ils sont uniquement composés d'énoncés correctement prononcés (phrases et mots isolés). Le texte demandé à l'élève a été phonétisé avec un dictionnaire de prononciation. L'ensemble de test contient 0,48 heures d'énoncés qui peuvent ou non contenir des erreurs de lecture. Dans cette étude, le test est uniquement composé de phrases, car elles contiennent une plus grande variété d'erreurs de lecture par rapport aux mots isolés. Les phonèmes lus par les enfants ont été transcrits manuellement par deux annotateur·rice·s. Les enregistrements ont été écartés en cas de désaccord.

Les erreurs de lecture faites par les enfants débutants sont diverses et parfois uniques, ce qui rend l'annotation manuelle des données assez difficile pour les experts humains, et peut causer d'autant plus de confusion pour un système automatique. Comprendre les différents types d'erreurs de lecture peut aider à analyser le comportement du système lorsqu'il les rencontre. Les pourcentages affichés représentent la proportion de chaque catégorie d'erreurs dans notre ensemble de test, ce qui fait un total de 13,1% de mots qui ne sont pas correctement lus.

— **Substitution** (5,1%) : un mot contient une ou plusieurs substitutions, insertions ou suppressions de phonèmes. Le mot résultant peut exister (environ 50% des cas), ou non ;
— **Répétition** (4,5%) : un mot est répété une ou plusieurs fois. Un mot répété peut également être substitué (contenir une erreur de déchiffrage, 23% des cas) ;
— **Suppression** (2,9%) : un mot est sauté par l'élève ;
— **Hésitation** (0.6%) : un mot contient un ou plusieurs silences dus à l'hésitation de l'enfant.

3 Description du système

Notre système de reconnaissance automatique de phonèmes est fondé sur l'architecture Transformer, entraîné sur de la parole adulte puis adapté sur une petite quantité de parole enfant (*Transfer Learning*). Il a été implémenté par nos soins en Pytorch.

3.1 Présentation du modèle Transformer

Proposé par (Vaswani *et al.*, 2017) et adapté à la reconnaissance automatique de la parole par (Dong *et al.*, 2018), le modèle Transformer suit une architecture *end-to-end* encodeur-décodeur séquence à séquence (*seq2seq*). Il se fonde uniquement sur des mécanismes d'attentions, abandonnant les réseaux de neurones récurrents habituels des systèmes *seq2seq*. La récurrence, essentielle pour extraire l'information de position des trames audio, est remplacée par des encodages positionnels qui sont concaténés aux encodages d'entrée initiaux, ainsi que par des modules d'auto-attention à plusieurs têtes et des réseaux de neurones à propagation avant tenant compte de la position. Le choix de cette architecture repose sur ses excellentes performances dans des tâches de reconnaissance de parole d'adulte (Karita *et al.*, 2019b), que nous avons confirmées sur la parole d'enfants apprenant·e·s lecteur·rice·s dans (Gelin *et al.*, 2021a). De par la nature des architectures seq2seq, qui donne les séquences de phonèmes de référence au décodeur durant l'apprentissage, ce dernier apprend un

modèle de langage de façon implicite, qui modélise un certain contexte linguistique.

L'architecture du Transformer utilisé dans ce travail suit celle de l'article original (Vaswani *et al.*, 2017). La procédure d'entraînement, ainsi que les hyper-paramètres utilisés, sont les mêmes que dans une étude antérieure (Gelin *et al.*, 2021a).

3.2 Transformer+CTC

Une fonction *Connectionist Temporal Classification* (CTC) (Graves *et al.*, 2006) peut être ajoutée au système Transformer en sortie de l'encodeur. Cette fonction a rendu possible l'entraînement des modèles de RAP *end-to-end* avec un simple encodeur en s'affranchissant de l'étape d'alignement explicite entre les séquences d'entrée et de sortie. Cette addition au système Transformer permet d'effectuer un entraînement multi-objectif CE+CTC avec la fonction initiale d'entropie croisée (CE pour *Cross-Entropy)*), et fournit la possibilité d'effectuer un décodage joint des sorties individuelles de l'encodeur et du décodeur.

3.2.1 *Entraînement multi-objectif CE+CTC*

L'entraînement multi-objectif peut être effectué avec les fonctions CE et CTC pour les systèmes encodeur-décodeur fondés sur l'attention tels que le Transformer (Watanabe *et al.*, 2017; Karita *et al.*, 2019a). Dans ce cadre, l'utilisation de l'objectif CTC vise à contraindre le mécanisme d'attention à trouver des alignements monotones. Cette contrainte est particulièrement intéressante pour reconnaître les répétitions de mots des apprenant·e·s lecteur·rice·s, car les mécanismes d'attention auront tendance à fusionner plusieurs occurrences d'un mot en une seule, manquant ainsi les mots répétés. La fonction CTC est combinée avec la fonction CE avec un certain poids (0,3 dans cette étude).

3.2.2 *Décodage joint CTC/attention*

Le Transformer+CTC a la particularité d'avoir deux sorties distinctes : une prédiction est faite par le décodeur grâce aux mécanismes d'attention, et une autre est faite par l'encodeur via la fonction CTC. Fondées sur des paradigmes opposés, ces sorties émettent des prédictions différentes. Afin de tirer parti des capacités de chacune sans avoir à choisir au cas par cas (impossible en pratique), le décodage joint CTC/attention a été proposé dans (Watanabe *et al.*, 2017). Les auteurs obtiennent des gains significatifs avec cette méthode par rapport à l'utilisation d'un algorithme *beam search* sur l'unique sortie du décodeur. Le décodage joint a été appliqué au Transformer dans (Karita *et al.*, 2019a).

3.3 Apprentissage par transfert

Des expériences préliminaires où nous avons essayé d'entraîner un Transformer uniquement sur les 13 heures de parole d'enfants que nous avons à disposition, ont mené à des taux de PER prohibitifs, supérieurs à 50%. Pour améliorer cet état de fait, l'approche naturelle est d'adapter un modèle entraîné sur une quantité bien plus grande de parole d'adultes, en se servant de ces 13 heures comme données d'adaptation. Il s'agit de la méthode d'apprentissage par transfert (*Transfer Learning*, TL), très

utilisée en apprentissage profond en général (Abad *et al.*, 2020; Duan *et al.*, 2020). Nous suivons les recommandations de (Shivakumar & Georgiou, 2020), où les auteurs suggèrent, pour de très jeunes enfants (5-8 ans), d'appliquer le TL sur l'ensemble des couches du modèle source.

4 Techniques d'augmentation

4.1 Augmentation de bruit de brouhaha

Le bruit de brouhaha (*babble noise* ou encore *cocktail party noise*) désigne le bruit engendré par un grand nombre de personnes parlant en même temps. Lorsque l'application Lalilo est utilisée en salle de classe, les enfants sont en autonomie, ce qui engendre inévitablement un niveau important de bruit de brouhaha sur certains enregistrements. L'entraînement multi-conditions (Rajnoha, 2009) est une technique d'augmentation de données qui consiste à superposer différents types de bruit, à différents niveaux de bruit, aux enregistrements de parole du jeu d'entraînement. Cela permet au modèle acoustique d'apprendre à reconnaître la parole dans le bruit environnant lors de son entraînement.

Notre corpus de bruit, baptisé *Lali-noise*, a été constitué d'enregistrements obtenus par la plateforme Lalilo ayant été manuellement étiquetés comme « bruités ». Ce corpus est représentatif des environnements de salles de classe et inclut des enfants qui parlent, des enseignant·e·s qui font cours, mais également des bruits typiques des salles de classe. Nous augmentons chaque enregistrement d'entraînement de façon à obtenir différentes versions avec différentes valeurs de rapport signal à bruit (RSB) : 2, 5, 10 et 15 dB. Le corpus augmenté contient ainsi la version originale et jusqu'à quatre versions augmentées, en fonction du RSB initial : un enregistrement avec un RSB initial de 8 dB ne sera augmenté que deux fois, à 2 et 5 dB.

4.2 Augmentation d'erreurs de lecture synthétiques

Les données reçues via la plateforme Lalilo contiennent des erreurs de lecture. Or notre jeu d'entraînement ne contient que des enregistrements de lecture correcte, car les erreurs sont coûteuses à annoter. Notre modèle n'est donc pas préparé à gérer les erreurs de lecture des enfants. Nous montrons ici comment créer des erreurs de lecture synthétiques à intégrer dans le jeu d'entraînement, dans l'optique d'améliorer la robustesse des modèles à la parole d'apprenant·e·s lecteur·rice·s. Les erreurs de prononciation et les répétitions étant les plus fréquentes (respectivement 5,1% et 4,5%), nous proposons de simuler ces deux types d'erreurs. Nous présumons que l'augmentation par simulation d'erreurs agira sur deux leviers : (1) apprendre aux mécanismes d'attention à distinguer les mots répétés malgré leur grande flexibilité, en leur présentant des exemples de répétitions de mots dans le corpus d'entraînement, (2) limiter la tendance du modèle de langage implicite du décodeur à gommer les erreurs de lecture de l'enfant, en lui faisant apprendre des séquences de phones plus diverses contenant des substitutions de mots et de phones.

Chaque enregistrement original d'entraînement est aligné en amont avec un GMM-HMM au niveau du mot et un TDNNF-HMM au niveau du phonème (Gelin *et al.*, 2021a), afin d'obtenir les limites temporelles de chaque mot et phonème prononcé par l'enfant. Nous utilisons trois types d'opérations :

— Répétition de mot(s) : les segments audio correspondant aux mots à répéter sont extraits et insérés dans l'enregistrement original, soit en début d'enregistrement, soit avant le mot initial ;
— Substitution de mot : un mot ressemblant au mot à substituer est choisi parmi une liste établie en amont (par exemple, le mot « roule » peut être substitué par « foule » ou « roula » en

modifiant un phonème). Un segment audio correspondant à ce mot de substitution est extrait d'un autre enregistrement, et inséré à la place du mot initial dans l'enregistrement original ;
— Substitution de phonème : contrairement à l'opération précédente, cette opération vise à simuler des mots n'existant pas dans la langue française. Nous descendons ainsi au niveau du phonème, en substituant un phonème par un autre de la même famille (voyelles, plosives, fricatives...) se trouvant dans le même enregistrement.

Les trois méthodes sont appliquées sur un sous-ensemble d'enregistrements (phrases et mots isolés) du jeu d'entraînement. Des sous-ensembles de données pour chaque type d'augmentation sont créés, puis regroupés pour correspondre à des proportions d'erreurs entre 1 et 10% (27 combinaisons testées). Cette fourchette a été choisie sur la base des proportions observées dans nos données. Les proportions qui ont donné la meilleure valeur de coût sur notre ensemble de validation ont été sélectionnées : 6,0% de répétitions de mots, 1,5% et 3,5% de substitution de mots et de phonèmes.

5 Évaluation et analyse des résultats

Dans cette étude, nous visons à transcrire avec précision ce que l'enfant a lu, y compris les éventuelles erreurs de lecture au niveau du phonème. Par conséquent, nous mesurons la performance avec un taux d'erreur phonème (PER), qui est calculé comme le nombre d'erreurs de reconnaissance de phonème sur le nombre de phonèmes de référence.

5.1 Ajout de la fonction CTC et impact du décodage joint

Le modèle Transformer source, entraîné sur environ 150 heures de parole d'adulte, a obtenu un PER de 7,5% sur un jeu de test de 7,2 heures de parole d'adulte Common Voice, mais un taux très élevé sur la parole d'enfant (44,6%). L'apprentissage par transfert permet d'atteindre un taux de 22,9%, comme indiqué dans le tableau 1. Ce système a surpassé un TDNNF-HMM hybride entraîné avec TL de 6,1% PER dans une étude précédente (Gelin *et al.*, 2021a).

Système	PER
Transformer	22,9
Transformer+CTC encodeur	26,6
Transformer+CTC décodeur	22,9
Transformer+CTC joint	**19,6**

TABLE 1 – PER (%) obtenus par les systèmes Transformer, Transformer+CTC (sorties individuelles encodeur et décodeur, et une sortie combinée)

Nous voyons dans le tableau 1 que l'entraînement multi-objectif seul n'apporte pas d'amélioration significative du PER : la sortie décodeur du Transformer+CTC obtient le même PER que le Transformer. Cette sortie, sur laquelle influe le modèle de langage implicite appris par le décodeur, obtient un taux d'erreurs significativement plus petit que la sortie encodeur. Ce résultat suggère que l'effet du modèle de langage implicite est globalement bénéfique pour la reconnaissance de phonèmes dans des phrases lues. La technique de décodage joint CTC/attention démontre une grande efficacité, puisqu'elle permet d'atteindre un PER de 19,6%, ce qui représente une amélioration absolue de 7,0% et de 3,3% par rapport aux sorties individuelles encodeur et décodeur, respectivement.

5.2 Augmentation de données

Le tableau 2 présente les taux PER obtenus sur le jeu de test Lalilo sans et avec les augmentations de données présentées précédemment. Nous voyons que l'augmentation de bruit *Lali-noise* réduit le PER de 2,5%. Nous observons également une légère amélioration (0,9%) du PER avec l'utilisation de la technique d'augmentation par simulation d'erreurs de lecture.

Augmentation	PER
Sans augmentation	19,6
Lali-noise-aug	17,1
ErrSyn-aug	18,7
ErrSyn-aug + *Lali-noise*-aug	**15,1**

TABLE 2 – PER (%) obtenus par le Transformer+CTC TL, sans et avec augmentations de données

Enfin, nous constatons que les deux techniques d'augmentation proposées dans cet article sont complémentaires : les combiner permet d'obtenir un PER de 15,1%, représentant une amélioration absolue de 4,5% par rapport au système sans augmentation. Il est intéressant de noter que la réduction de PER apportée par la technique ErrSyn-aug est supérieure (2% absolus) lorsque combinée avec l'augmentation de bruit que lorsque seule (0,9%). Nous suggérons que ce phénomène est lié au fait que les versions originale et ErrSyn-aug de chaque enregistrement sont augmentées avec deux enregistrements de bruit différents : cela implique une plus grande diversité des données dans le jeu d'entraînement.

5.3 Amélioration de la robustesse au bruit

La figure 1 présente les performances de ces modèles en fonction du niveau de bruit grâce à trois intervalles de RSB : RSB $\leq$ 10 dB (très bruité), 10 dB $<$ RSB $\leq$ 20 dB (moyennement bruité) et RSB $>$ 20 dB (non bruité). Nous observons tout d'abord que la présence de bruit a un fort impact sur les performances du Transformer+CTC : une perte relative d'environ 30% de PER se remarque entre chaque niveau de bruit, pour le modèle sans augmentation. Nous voyons également que l'augmentation de bruit remplit sa fonction : le modèle augmenté obtient de meilleures performances sur les enregistrement très bruités (réduction de 5,3%) et moyennement bruités (4,0%).

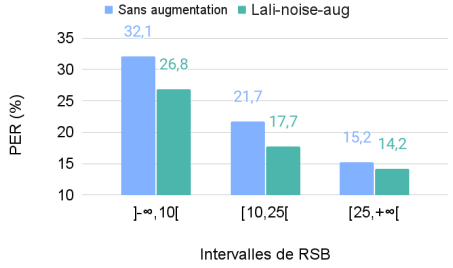

FIGURE 1 – PER (%) du Transformer+CTC avec et sans augmentation *Lali-noise*, pour différents intervalles de RSB

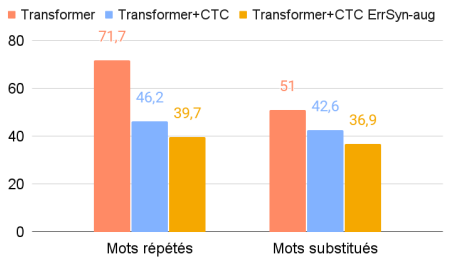

FIGURE 2 – PER (%) du Transformer, Transformer+CTC et Transformer+CTC ErrSyn-aug pour les mots répétés ou substitués

5.4 Amélioration de la robustesse aux erreurs de lecture

Comme mentionné dans la section 3.2.1, les mécanismes d'attention sont très flexibles, ce qui leur fait manquer certaines disfluences typiques des lecteur·rice·s débutant·e·s, telles que les répétitions. De plus, le décodeur est influencé par le modèle de langage implicite appris grâce au contexte linguistique vu en entraînement, couvrant les erreurs des élèves en reconnaissant ce qui aurait dû être lu, et non ce qui a réellement été lu. Nous proposons une étude sur l'efficacité de la fonction CTC et de la synthèse d'erreurs pour améliorer la robustesse du système sur les erreurs de lecture.

La figure 2 détaille le PER obtenu uniquement sur les mots répétés ou substitués. Nous voyons que le PER obtenu par le Transformer sur les mots répétés est extrêmement haut, ce qui est dû à la trop grande flexibilité des mécanismes d'attention. L'ajout de la fonction CTC améliore drastiquement ce taux : l'entraînement multi-objectif CE+CTC apporte une contrainte de monotonie aux mécanismes d'attention, qui leur permet de mieux détecter les répétitions. De plus, le décodage joint met à profit la grande efficacité de la sortie encodeur sur les mots répétés. Enfin, l'augmentation ErrSyn réduit de 6,5% supplémentaires le PER sur les mots répétés : la simulation de répétitions dans le jeu d'entraînement permet aux mécanismes d'attention d'apprendre à les gérer.

Nous voyons que le Transformer+CTC, grâce à l'entraînement CE+CTC et le décodage joint CTC/attention, surpasse également le Transformer sur les mots substitués. Le Transformer+CTC tire parti de la flexibilité de l'attention de la sortie décodeur tout en limitant l'effet du modèle de langage implicite du décodeur, qui a tendance à gommer les erreurs de lecture, par la pondération avec la sortie encodeur. L'augmentation ErrSyn réduit, là aussi, le PER de 5,7% : la simulation de substitutions de mots, par des mots existants ou non, augmente la diversité du contenu d'apprentissage. Cela permet au modèle de langage implicite d'incorporer du contexte linguistique correspondant à des erreurs de lecture, et ainsi de limiter sa tendance à ne reconnaître que le contenu à lire.

6 Conclusion

Cet article porte sur la reconnaissance de phonèmes dans la parole d'enfants apprenant·e·s lecteur·rice·s en situation de salle de classe. En plus des caractéristiques complexes acoustiques et prosodiques de la parole d'enfants, notre système doit gérer la présence de bruit de brouhaha et d'erreurs de lecture. Nous proposons ici différentes méthodes pour adapter un système de RAP fondé sur une architecture *end-to-end* Transformer à la parole de jeunes lecteur·rice·s. Nous utilisons l'apprentissage par transfert (TL) pour obtenir un PER de base de 22,9%, qui surpasse un modèle DNN-HMM adapté par TL de 6,1%. L'apprentissage multi-objectif avec CTC et le décodage joint CTC/attention améliorent les performances de 3,3% absolus, atteignant un PER de 19,6%. Nous proposons ensuite d'améliorer la robustesse de notre système au bruit de salle de classe et aux erreurs de lecture grâce à deux augmentations, l'une par ajout de bruit et l'autre par simulation d'erreurs de lecture. La combinaison de ces deux augmentations permet d'atteindre un PER de 15,1%. Nous analysons enfin le comportement du système quant à ces deux défis, et montrons que l'augmentation de bruit de salle de classe est efficace pour améliorer la précision du système sur les enregistrements bruités. L'ajout de la fonction CTC et l'augmentation d'erreurs synthétiques réduisent de plus drastiquement le PER sur les mots contenant des erreurs de lecture.

Les travaux futurs viseront à améliorer encore la performance sur les erreurs de lecture, en remplaçant l'attention par une attention monotone et localisée (Chorowski *et al.*, 2015; Tjandra *et al.*, 2017).

Références

ABAD A., BELL P., CARMANTINI A. & RENALS S. (2020). Cross lingual transfer learning for zero-resource domain adaptation. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 6909–6913.

BOLAÑOS D., COLE R., WARD W., BORTS E. & SVIRSKY E. (2011). FLORA : Fluent oral reading assessment of children's speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, **7**(4), 16.

CHOROWSKI J., BAHDANAU D., SERDYUK D., CHO K. & BENGIO Y. (2015). Attention-based models for speech recognition. In *Proc. of the International Conference on Neural Information Processing Systems (NIPS)*, p. 577–585 : MIT Press.

DONG L., XU S. & XU B. (2018). Speech-transformer : A no-recurrence sequence-to-sequence model for speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 5884–5888.

DUAN R., KAWAHARA T., DANTSUJI M. & NANJO H. (2020). Cross-lingual transfer learning of non-native acoustic modeling for pronunciation error detection and diagnosis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, **28**, 391–401.

FRINGI E., LEHMAN J. F. & RUSSELL M. J. (2015). Evidence of phonological processes in automatic recognition of children's speech. In *Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Dresden*, p. 1621–1624.

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

GELIN L., PELLEGRINI T., PINQUIER J. & DANIEL M. (2021b). Simulating Reading Mistakes for Child Speech Transformer-Based Phone Recognition. In *Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Brno*, p. 3860–3864.

GODDE E., BAILLY G., ESCUDERO D., BOSSE M.-L. & ESTELLE G. (2017). Evaluation of reading performance of primary school children : Objective measurements vs. subjective ratings. In *Proc. of the International Workshop on Child Computer Interaction (WOCCI)*, p. 23–27.

GRAVES A., FERNÁNDEZ S., GOMEZ F. & SCHMIDHUBER J. (2006). Connectionist temporal classification : labelling unsegmented sequence data with recurrent neural networks. In *Proc. of the International Conference on Machine learning (ICML)*, p. 369–376.

KARITA S., SOPLIN N. E. Y., WATANABE S., DELCROIX M., OGAWA A. & NAKATANI T. (2019a). Improving Transformer-Based End-to-End Speech Recognition with Connectionist Temporal Classification and Language Model Integration. In *Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Graz*, p. 1408–1412.

KARITA S., WANG X., WATANABE S., YOSHIMURA T., ZHANG W., CHEN N., HAYASHI T., HORI T., INAGUMA H., JIANG Z., SOMEKI M., SOPLIN N. E. Y. & YAMAMOTO R. (2019b). A Comparative Study on Transformer vs RNN in Speech Applications. *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, (April 2020), 449–456.

LEE S., POTAMIANOS A. & NARAYANAN S. S. Y. (1999). Acoustics of children's speech : developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, **105**(3), 1455–1468.

MOSTOW J. & AIST G. (2001). Evaluating tutors that listen : An overview of Project LISTEN. In *Smart machines in education : The coming revolution in educational technology.*, p. 169–234. The MIT Press.

MUGITANI R. & HIROYA S. (2012). Development of vocal tract and acoustic features in children. *The Journal of the Acoustical Society of Japan*, **68**(5), 234–240.

POTAMIANOS A. & NARAYANAN S. (1998). Spoken dialog systems for children. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, p. 197–200 vol.1.

POTAMIANOS A. & NARAYANAN S. (2003). Robust Recognition of Children's Speech. *IEEE Transactions on Speech and Audio Processing*, **11**(November 2003), 603–616.

RAJNOHA J. (2009). Multi-condition training for unknown environment adaptation in robust asr under real conditions. *Acta Polytechnica*, **49**.

SERIZEL R. & GIULIANI D. (2014). Deep neural network adaptation for children's and adults' speech recognition. In *Proc. of the Italian Computational Linguistics Conference (CLiC-it)*, p. 137–140.

SHIVAKUMAR P. G. & GEORGIOU P. (2020). Transfer learning from adult to children for speech recognition : Evaluation, analysis and recommendations. *Computer Speech & Language*, **63**, 101077.

TJANDRA A., SAKTI S. & NAKAMURA S. (2017). Local Monotonic Attention Mechanism for End-to-End Speech and Language Processing. *ArXiv preprint :1705.08091*.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER U. & POLOSUKHIN I. (2017). Attention is all you need. In *Proc. of the International Conference on Neural Information Processing Systems (NIPS)*, p. 6000–6010, Red Hook, NY, USA : Curran Associates Inc.

WATANABE S., HORI T., KIM S., HERSHEY J. R. & HAYASHI T. (2017). Hybrid CTC/Attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, **11**(8), 1240–1253.

YEUNG G. & ALWAN A. (2018). On the difficulties of automatic speech recognition for kindergarten-aged children. In *Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Hyderabad*, p. 1661–1665.