Reconnaissance de phones fondée sur du Transfer Learning pour des enfants apprenants lecteurs en environnement de classe

Lucile Gelin, Morgane Daniel, Thomas Pellegrini, Julien Pinquier

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Reconnaissance de phones fondée sur du Transfer Learning pour des enfants apprenants lecteurs en environnement de classe

Lucile Gelin[1, 2] Morgane Daniel[2] Thomas Pellegrini[1] Julien Pinquier[1]
(1) IRIT, Université Paul Sabatier, CNRS, Toulouse, France
(2) Lalilo, Paris, France
lucile.gelin@irit.fr, morgane@lalilo.com

RÉSUMÉ

A conditions égales, les performances actuelles de la reconnaissance vocale pour enfants sont inférieures à celles des systèmes pour adultes. La parole des jeunes enfants est particulièrement difficile à reconnaître, et les données disponibles sont rares. En outre, pour notre application d'assistant de lecture pour les enfants de 5-7 ans, les modèles doivent s'adapter à une lecture lente, des disfluences et du bruit de brouhaha typique d'une classe. Nous comparons ici plusieurs modèles acoustiques pour la reconnaissance de phones sur de la parole lue d'enfant avec des données bruitées et en quantité limitée. Nous montrons que faire du *Transfer Learning* avec des modèles entraînés sur la parole d'adulte et trois heures de parole d'enfant améliore le taux d'erreur au niveau du phone (PER) de 7,6% relatifs, par rapport à un modèle enfant. La normalisation de la longueur du conduit vocal sur la parole d'adulte réduit ce taux d'erreur de 5,1% relatifs supplémentaires, atteignant un PER de 37,1%.

ABSTRACT

Transfer Learning based phone recognition on children learning to read, with speech recorded in a classroom environment

Current performance of speech recognition for children is below that of the state-of-the-art for adult speech. Young children's speech is particularly difficult to recognise, and substantial corpora are missing to train acoustic models. Furthermore, in the scope of our reading assistant for 5-7-year-old children learning to read, models need to cope with slow reading rate, disfluencies, and classroom-typical babble noise. In this paper, we compare acoustic models for phone recognition on child speech using data that is very noisy and limited in quantity. We show that transfer learning with adult-trained time-delay neural networks and three hours of child speech improves the phone error rate by 7.6% relative, over a model trained on child speech. The addition of vocal tract length normalisation on adult speech further reduces the error rate by 5.1% relative, reaching a PER of 37.1%.

MOTS-CLÉS : Reconnaissance de phones, parole d'enfant, apprentissage par transfert, normalisation de la longueur du conduit vocal, réseau de neurones à délai temporel.

KEYWORDS: Phone recognition, child speech, transfer learning, vocal tract length normalisation, time-delay neural network.

1 Introduction

Le corps humain, et en particulier l'appareil de production de la parole, évolue continuellement pendant les premières années de la vie. Entre 5 et 7 ans, les mécanismes articulatoires des enfants ne sont pas stables, ce qui implique une variabilité spectrale intra- et inter-locuteurs. La stabilisation

Volume 1 : Journées d'Études sur la Parole, pages 253–261.
HAL : hal-02798545.

du contrôle vocal ne se produit que vers l'âge de 8 ans (Lee *et al.*, 1999). En raison de la croissance lente du conduit vocal, leurs fréquence fondamentale et formants n'atteignent les niveaux d'un adulte mature qu'à l'âge de 15 ans (Mugitani & Hiroya, 2012). En outre, les erreurs phonologiques, comme la suppression de syllabes faibles ou la substitution de phones due à un mauvais positionnement de la langue et des lèvres (Fringi *et al.*, 2015), sont très courantes dans la parole des jeunes enfants, et ont tendance à disparaître avec l'âge. 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; Proença, 2018). Le projet FLUENCE notamment travaille depuis quelques années sur l'évaluation automatique de la fluence auprès des apprenants lecteurs français (Godde *et al.*, 2017). Travailler sur la parole des lecteurs non-experts ajoute des difficultés dues à la présence de nombreuses disfluences.

Lalilo [1] propose un assistant de lecture pour les enfants de 5 à 7 ans, avec un exercice de lecture à voix haute. Pour cela, nous entraînons un classifieur qui donne une décision binaire sur la lecture correcte ou non du mot. Cet article présente nos travaux sur la modélisation acoustique, avec pour objectif l'amélioration de la précision de la reconnaissance automatique de phones, dont découle la précision du classifieur.

Dans la section 2, nous présentons nos motivations et les techniques utilisées. Le dispositif expérimental est détaillé dans la section 3, suivi des résultats dans la section 4. Enfin, nous analysons le comportement de nos modèles en fonction du voisement, et en présence d'erreurs de lecture.

2 Méthodes

En raison du peu de données de parole d'enfant disponibles, les systèmes fondés sur les réseaux neuronaux profonds n'ont commencé à être exploités que récemment pour la reconnaissance automatique de la parole d'enfant. Les architectures hybrides Deep Neural Network - Hidden Markov Model (DNN-HMM) de modèles acoustiques ont été largement utilisées dans la reconnaissance vocale au cours des deux dernières décennies. (Serizel & Giuliani, 2014a) utilisent un système DNN-HMM, entraîné conjointement sur des données adultes et enfants, puis adapté à la parole d'un groupe d'âge spécifique. Pour une application d'apprentissage d'une seconde langue chez les enfants, (Metallinou & Cheng, 2014) ont présenté un DNN-HMM qui, même entraîné sur très peu de données, a surpassé des systèmes basés sur des mélanges de lois gaussiennes (GMM-HMM).

Dans cet article, nous utilisons un système hybride où le DNN est un Time-Delay Neural Network (TDNN), architecture introduite par (Waibel *et al.*, 1989) pour la reconnaissance de phones. Ce type de réseau s'est révélé particulièrement adapté à la RAP, de par sa capacité à représenter les relations entre événements acoustiques dans le temps, ainsi qu'à fournir une invariance temporelle des paramètres appris par le réseau (Waibel *et al.*, 1989). Pour cela, la largeur du contexte varie selon les couches : les couches inférieures apprennent des caractéristiques acoustico-phonétiques de courte durée, tandis que les couches supérieures apprennent des caractéristiques plus complexes de plus longue durée. Les TDNN ont été utilisés avec succès pour la reconnaissance des voyelles sur la parole des enfants dans une langue peu dotée (Yong & Ting, 2011).

Le Transfer Learning (TL) permet de surmonter le manque de données dans un domaine spécifique : cette méthode consiste à transférer des connaissances préalables acquises sur un grand corpus hors

1. https://www.lalilo.com/

domaine vers un modèle acoustique entraîné sur un petit corpus correspondant à l'application. Le système bénéficie ainsi de connaissances de bases qui peuvent être adaptées à un domaine ou à une tâche spécifique. Cette méthode a montré une amélioration des performances dans des applications de transfert de langues (Shi *et al.*, 2018) et de transfert d'âge (Shivakumar & Georgiou, 2018).

Nous explorons ici l'adaptation de modèles acoustiques en utilisant un modèle TDNN entraîné sur un grand corpus de parole adulte, et deux méthodes de TL. La première consiste à ré-initialiser aléatoirement les deux dernières couches du réseau, et à les entraîner avec des données d'enfant, ce qui a pour objectif d'adapter fortement le modèle à la parole d'enfant. La seconde prend toutes les couches existantes du modèle source et les ré-entraîne avec des données d'enfant, et ainsi garde toutes les informations apprises par le modèle sur les voix d'adulte. Nous utilisons différents facteurs d'apprentissage pour équilibrer les connaissances pré-acquises sur la parole d'adulte et les caractéristiques acoustiques nouvellement acquises sur la parole d'enfant.

Une adaptation approfondie est réalisée avec la technique VTLN. Tandis que Serizel et al. (Serizel & Giuliani, 2014b) utilisent la VTLN sur un corpus mixte enfants-adultes en normalisant les caractéristiques de chaque locuteur, nous proposons d'utiliser la VTLN pour étendre les fréquences d'adulte vers des fréquences d'enfant et entraîner un TDNN adulte sur ces paramètres transformés. La VTLN n'est donc pas appliqué de façon individuelle à chaque locuteur, mais de façon globale à tous les locuteurs avec des facteurs de déformation fixes : un pour les femmes et un pour les hommes. Cela nous permet d'obtenir une gamme de fréquences proche de celle des enfants pour chaque locuteur, minimisant ainsi la différence entre parole d'adulte et parole d'enfant, avec pour objectif une meilleure efficacité du transfer learning. La principale contribution réside dans l'utilisation de ce TDNN adulte adapté aux voix d'enfants avec de la VTLN comme modèle source pour le transfer learning.

3 Dispositif expérimental

3.1 Données de parole

Nous utilisons deux jeux de données de parole en français : le corpus adulte *Commonvoice* [2], et un corpus enfant interne, appelé *Lalilo* par la suite. Le tableau 1 présente des informations sur ces deux corpus.

TABLE 1 – Informations sur les données de parole

Corpus	**Commonvoice**		**Lalilo**		
Set	Train	Test	Train	Test C	Test I
Nb locuteurs	78	268	562	69	153
Durée (h)	20,0	9,0	3,8	0,4	0,4
Durée moyenne (s)					
Par enregistrement	3,6	3,8	7,0	2,3	3,4
Par locuteur	127,4	-	22,0	-	-
RSB moyen (dB)	35,3 ± 16,1	32,9 ±14,2	25,6 ± 13,9	23,9 ±11,6	20,5 ± 11,9

Le jeu de données Commonvoice est composé de phrases lues par des adultes, tâche qui se rapproche de la lecture à voix haute des enfants. 72% des locuteurs sont masculins, et 7% féminin, les autres locuteurs n'ayant pas fourni cette information. Chaque enregistrement a été validé par 2 ou 3 annotateurs, le corpus ne contient donc que peu d'erreurs. En outre, il présente un rapport signal à bruit (RSB) moyen élevé.

2. Corpus disponible : https://voice.mozilla.org/fr

Le corpus Lalilo contient des enregistrements d'enfants de la grande section au CE1, âgés de 5 à 7 ans, lisant à haute voix des mots isolés, des phrases et des histoires courtes. Les enregistrements ont été recueillis soit directement dans les écoles, soit par le biais d'un exercice de lecture à voix haute sur la plateforme Lalilo. Dans le premier cas, les conditions environnementales sont relativement propres : un microphone de bonne qualité est utilisé, et le niveau de bruit est contrôlé. Dans le second cas, cependant, les enseignants laissent généralement un petit groupe d'élèves travailler en autonomie sur la plateforme, ce qui implique inévitablement la présence de bruit de brouhaha sur les enregistrements. Le tableau 1 affiche la moyenne et l'écart-type du RSB pour chaque ensemble de données. Par rapport au corpus Commonvoice, le RSB moyen est significativement plus faible.

Les données d'apprentissage ne contiennent que des histoires, phrases ou mots correctement prononcés et lus avec fluidité. En accord avec la tâche de l'assistant numérique qui vise à détecter les erreurs de déchiffrage et de fluence sur des mots isolés, l'ensemble de test est formé uniquement de mots isolés [3], la durée moyenne des enregistrements est donc plus faible que celle du corpus d'apprentissage. Les phones réellement lus par les élèves ont été transcrits par deux juges humains. Chaque mot a également été classé entre trois catégories :

— Correct : lecture correcte et fluide.
— Erreur de fluence (*Fluence*) : lecture correcte mais non fluide (hésitations, faux départs...)
— Erreur de déchiffrage (*Déchiffrage*) : lecture incorrecte, avec au moins un phone mal lu.

Deux sous-ensembles de test ont été créés à partir de ces catégories : Le test C contient les mots bien lus, et le test I contient les mots qui comportent des erreurs de lecture, c'est-à-dire de fluence ou de déchiffrage. Dans la catégorie *Déchiffrage*, les phones peuvent être soit substitués, soit supprimés, soit insérés. Les erreurs de déchiffrage prévalent sur les erreurs de fluence, car les premières sont plus répréhensibles que les secondes lors de l'évaluation du niveau de lecture d'un enfant. Ainsi, les mots classés comme erreurs de déchiffrage peuvent également contenir des erreurs de fluence.

3.2 Système de reconnaissance de phones

Toutes les expériences sont réalisées avec l'outil Kaldi (Povey *et al.*, 2011).

3.2.1 Paramètres acoustiques

Pour les modèles GMM-HMM servant à la génération des alignements pour l'entraînement des TDNN, les paramètres sont des Mel-frequency cepstral coefficients (MFCC) de dimension 13 avec une fenêtre de 25 ms et un décalage de 10 ms, auxquels nous ajoutons des dérivées première et seconde. Les TDNN sont alimentés par des MFCC haute résolution de dimension 40. Nous effectuons également de l'augmentation de données en déformant temporellement le son brut avec des facteurs de 0,9, 1,0 et 1,1. Dans la plupart des systèmes de la litérature, des i-vecteurs sont utilisés pour augmenter les paramètres avec des informations spécifiques au locuteur. Il a en effet été observé que cela améliorait les performances de modèles entraînés et testés sur de la parole d'adultes. Cependant, que nous entraînions l'extracteur de i-vecteurs sur des données d'adultes ou d'enfants, les performances étaient toujours détériorées par rapport à celles des modèles sans i-vecteurs. Dans le premier cas, les caractéristiques extraites sur la base de la parole d'adulte ne correspondaient pas à la parole d'enfant. Dans le second cas, les informations extraites à partir de la parole d'enfant n'étaient pas pertinentes de par la faible quantité de données et la faible durée moyenne de parole par locuteur (voir tableau 1). Les résultats présentés dans la section 4 sont donc obtenus sans l'utilisation de i-vecteurs.

3. Exemples audio disponibles : https://frama.link/JEP2020-exemples-audio

3.2.2 Modèles chain-TDNN

Les chain-TDNN, appelés TDNN par la suite, ont été implémentés avec l'architecture présentée dans (Peddinti *et al.*, 2015), et en s'appuyant sur la recette Kaldi du corpus Commonvoice [4]. Nous utilisons dans cet article une architecture de modèle chaîne (Povey *et al.*, 2016), qui diffère d'un DNN-HMM classique par l'utilisation d'une fonction de coût au niveau de la séquence plutôt qu'au niveau de la trame. La procédure d'entraînement est similaire à un entraînement par maximisation de l'information mutuelle sans graphe de phones (Veselý *et al.*, 2013).

Une autre caractéristique du modèle chaîne est sa fréquence de trame divisée par 3 à la sortie du réseau : cela accélère le calcul et permet d'effectuer une augmentation de données en appliquant un décalage de trame de 0, 1 et 2 trames.

Les chain-TDNN présentés dans cet article sont similaires à ceux spécifiés dans (Povey *et al.*, 2016). Ils contiennent une première couche affine LDA prenant comme trames d'entrée les indices -2,-1,0,1,2. Suivent ensuite 9 couches cachées avec ReLU, chacune avec 768 unités, parmi lesquelles les couches 2, 4, 6, 7, 8 sont configurées avec des indices de concaténation -1,0,1 -1,0,1 -3,0,3 -3,0,3 -6,-3,0. La couche de sortie comporte 496 unités.

Le HMM utilisé dans notre système hybride TDNN-HMM est un modèle monophone, car nous avons observé de meilleurs résultats qu'avec des modèles triphones. En effet, notre corpus de parole d'enfant étant très réduit, l'entraînement souffre du faible nombre d'occurrence de chaque triphone. De plus, les erreurs produites par les enfants pourraient correspondre à des triphones non représentés dans la langue française, et donc ne pas être reconnues.

3.2.3 Modèles Transfer Learning

Nous avons exploré deux méthodes de TL [5]. La première méthode consiste à supprimer les deux dernières couches du modèle TDNN source et à les remplacer par deux couches initialisées de façon aléatoire. La ré-initialisation doit permettre au réseau de s'adapter rigoureusement aux caractéristiques des enfants. La seconde méthode garde les couches existantes du modèle source, conservant ainsi certaines informations acoustiques de la parole d'adulte. Dans les deux méthodes, les couches finales, ainsi que les autres couches transférées du modèle source sont ré-entraînées avec des facteurs d'apprentissage respectifs de 1 et 0,25. Choisir un facteur d'apprentissage de 0 pour les couches transférées revient à utiliser uniquement les connaissances pré-acquises sur la parole d'adulte. Pour ré-entraîner les modèles TDNN avec des données d'enfants, nous pouvons fournir des alignements soit à partir d'un modèle GMM-HMM, soit à partir d'un modèle TDNN.

3.2.4 Vocal Tract Length Normalisation

Nous appliquons la normalisation VTLN aux MFCC extraits du corpus d'entraînement Commonvoice afin d'étendre la gamme de fréquences des adultes vers une gamme proche de celle des enfants. Les facteurs de déformation qui minimisent le PER obtenus par les modèles GMM-HMM sur l'ensemble Lalilo Test C sont de 1,2 pour les femmes et de 1,3 pour les hommes.

4 Évaluation

Dans cette section, nous testons plusieurs modèles acoustiques enfant (Lalilo), adulte (Commonvoice), TL et TL + VTLN sur le jeu de test C de Lalilo, c'est-à-dire sur des mots qui ont été correctement lus

4. Disponible au lien : https://frama.link/script-TDNN
5. Inspirées par les recettes Kaldi disponibles au lien : https://frama.link/scripts-TL

par des enfants. Les résultats sont affichés dans le tableau 2.

Nous ne visons pas ici à reconstituer et à corriger les mots en fonction des phones détectés, mais à repérer les substitutions, insertions et suppressions de phones chez les enfants qui lisent à voix haute. Par conséquent, nous ne mesurons pas les performances avec un WER, comme le font la plupart des études de RAP, mais avec un taux d'erreurs sur les phones (Phone Error Rate, PER). Le PER est défini comme le ratio entre le nombre d'erreurs (insertions, substitutions et suppressions) et le nombre de phones de la prononciation de référence. Dans cette même optique, nous utilisons un modèle de language unigramme appris sur le corpus d'entraînement de Lalilo pour le décodage.

4.1 Modèles chain-TDNN

Nous validons notre système de reconnaissance de phones en entraînant et testant un modèle TDNN sur le corpus Commonvoice. Les résultats sont affichés dans le tableau 2 : on atteint un PER de 28,4%.

TABLE 2 – Caractéristiques des différents modèles acoustiques et PER (%) obtenus

Nom du modèle	**Méthode de TL**	**Alignements générés par**	**VTLN**	**PER (%)**	
				Commonvoice	**Lalilo Test C**
Commonvoice	–	GMM-HMM	Non	28,4	72,5
Commonvoice + VTLN	–	GMM-HMM	Oui	38,6	69,6
Lalilo	–	GMM-HMM	Non	–	42,3
TL 1	1	GMM-HMM	Non	–	44,0
TL 2A	2	GMM-HMM	Non	–	43,0
TL 2B	2	TDNN	Non	–	39,1
TL 2B + VTLN	2	TDNN	Oui	–	37,1

Le TDNN enfant (Lalilo), atteignant un PER de 42,3%, est 41,4% relatifs plus performant que le TDNN adulte dans la tâche de reconnaissance de phones sur de la parole d'enfant (PER de 72,5%), alors qu'entraîné sur 5 fois moins de données. L'utilisation de la VTLN sur la parole d'adulte permet d'améliorer le PER sur la parole d'enfant, au prix d'une dégradation sur la parole d'adulte.

4.2 Modèles Transfer Learning

Les deux méthodes détaillées en section 3.2.3 sont appelées TL 1 et TL 2 dans le tableau 2. Les alignements utilisés sont générés avec les modèles enfant car ils ont montré de meilleures performances que les modèles adultes. Dans le tableau 2, les noms A et B correspondent respectivement aux alignements générés par GMM-HMM et TDNNF.

Le modèle TL 1 donne de moins bons résultats que le modèle TDNN enfant (Lalilo). La méthode TL 2A donne des résultats légèrement meilleurs que l'approche TL 1 précédente, mais toujours sans amélioration par rapport au modèle Lalilo. La meilleure performance de l'approche TL 2 n'était pas attendue, puisque les auteurs de la méthode ont obtenu de meilleurs résultats avec la méthode TL 1, en utilisant le corpus Wall Street Journal comme source et le corpus Resource Managment (3 heures) comme cible. Cela est dû au fait qu'ils adaptent les tâches de reconnaissance (de la parole radio aux commandes vocales) alors que nous adaptons les domaines de reconnaissance : lors de la ré-initialisation des couches proches de la sortie, ils ne perdent que les informations liées à la tâche, qu'ils peuvent remplacer grâce aux données cible. Dans notre cas, nous perdons de précieuses

informations acoustiques et de prononciation qui ne peuvent pas être retrouvées avec uniquement quatre heures de parole d'enfant.

Enfin, la méthode TL 2B, qui utilise des alignements générés par un TDNN enfant, apporte une amélioration substantielle, avec un PER de 39,1 % pour le modèle TL 2B, ce qui correspond à une amélioration relative de 7,6 % par rapport au TDNN enfant.

4.3 Impact de la VTLN

La VTLN apportant une amélioration sur le TDNN Commonvoice lorsque testé sur parole d'enfant, une tendance identique est attendue pour le modèle TL 2B +VTLN, obtenu par transfer learning avec le modèle TDNN Commonvoice + VTLN. Le PER diminue, de 39,1% pour le modèle TL 2B, à 37,1% pour le modèle TL 2B VTLN, ce qui correspond à une amélioration relative de 5,1%.

5 Analyses & discussion

5.1 Analyse d'erreurs en fonction du voisement

Le taux d'erreur de reconnaissance (TER) est défini dans l'équation (1), avec C, S, D référant respectivement aux nombres de détections correctes, substitutions et suppressions. Il mesure la capacité du système à reconnaître correctement un phone donné, et donc ne prend pas en compte les insertions comme le PER.

$$TER = \frac{S + D}{C + S + D} \tag{1}$$

La figure 1 montre la capacité des différents modèles à reconnaître les phones voisés et non-voisés.

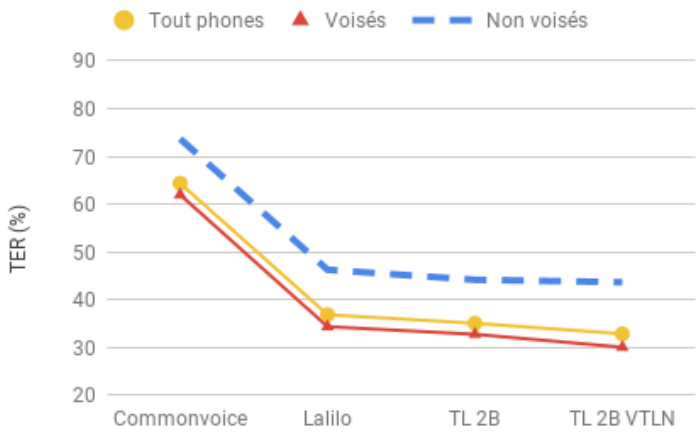

FIGURE 1 – Évolution du TER (%) sur les phones voisés et non-voisés du jeu de test C (Lalilo), pour les modèles adulte, enfant, TL 2B et TL 2B VTLN

La première observation est que les phones voisés sont, en moyenne, 24,6 % relativement mieux reconnus que les phones non-voisés. Les phones voisés, qui correspondent en français à 25 phones sur 31, représentent 79 % de notre corpus de test C, ce qui explique pourquoi la courbe "Tout phones" suit de si près celle des phones voisés. Le modèle enfant (Lalilo) réduit considérablement le TER par rapport au modèle adulte. De même, le modèle TL 2B améliore la reconnaissance des phones voisés et non voisés, avec des améliorations relatives respectives du TER de 4,7 % et 4,5 % par rapport au modèle enfant. Le dernier point, correspondant au modèle TL 2B VTLN, montre que la VTLN a une forte influence sur le TER pour les phones voisés (amélioration relative de 8,2%), mais n'apporte pas d'amélioration significative pour les phones non-voisés. Les phones non-voisés sont

en effet articulés sans aucune vibration des cordes vocales, ce qui signifie qu'ils ne possèdent pas de fréquence fondamentale ni d'harmoniques. Comme la VTLN agit sur les fréquences, il affecte les phones voisés qui sont caractérisés par leurs formants, mais pas les phones non-voisés. La très légère amélioration pourrait être due à la présence habituelle de pics de fréquence dans les consonnes, correspondant à la position des résonateurs, qui pourraient être modérément affectés par la VTLN.

5.2 Influence de la parole de lecteurs non-experts

Pour quantifier la difficulté apportée par des lecteurs non-experts par rapport à des lecteurs plus expérimentés dans le cadre d'un système de RAP, nous avons calculé le PER obtenu par le modèle monophone TL 2B VTLN sur les jeux de test C (lecture correcte) et I (lecture incorrecte) du corpus Lalilo, et obtenu les résultats affichés dans le tableau 3. Nous pouvons constater une détérioration relative drastique de 45,0% du PER pour le Test I, relativement au Test C.

TABLE 3 – PER (%) pour le modèle monophone TL 2B VTLN

Corpus	Test C	Test I (tous)	Test I (Fluence)	Test I (Déchiffrage)
PER (%)	37,1	53,8	51,1	56,9

Les catégories d'erreurs de lecture (Fluence et Déchiffrage) sont décrites en section 3.1. Nous observons que la catégorie Fluence obtient un PER relativement 10% meilleur que la catégorie Déchiffrage. En effet, les erreurs de déchiffrage prévalant sur les erreurs de fluence lors du classement des enregistrements, certains mots de la catégorie Déchiffrage contiennent les deux sortes d'erreurs, combinant ainsi les difficultés. Ces deux catégories correspondent à la réalité de notre application, et les résultats présentés démontrent la difficulté à reconnaître de la parole d'enfant apprenant lecteur.

6 Conclusion & perspectives

Dans le cadre d'une tâche de détection des erreurs pour l'évaluation de la lecture à haute voix chez les enfants, la précision du système de reconnaissance vocale a une grande incidence sur la pertinence des paramètres qui sont transmis à un classifieur. Dans cet article, nous améliorons la précision de la reconnaissance de phones sur parole d'enfant au moyen d'une méthode de *Transfer Learning*, où des modèles TDNN adultes sont adaptés avec quatre heures de parole d'enfant. Nous obtenons un PER de 39,1 %, ce qui correspond à un gain relatif de 7,6 % par rapport à un TDNN entraîné uniquement sur parole d'enfant. L'application de la VTLN sur le corpus d'entraînement de parole d'adulte pour l'extension de la gamme de fréquences de locuteurs, et l'utilisation de ce modèle adulte adapté au VTLN comme modèle source pour le transfer learning réduit le PER à 37,1%, ce qui apporte une réduction relative supplémentaire de 5,1%.

Nous utilisons actuellement une structure de TDNN factorisés qui ont montré de meilleurs résultats sur un petit corpus de parole d'enfant (Wu *et al.*, 2019). Nous pourrions également renforcer notre utilisation de la VTLN en utilisant un DNN pour adapter les facteurs de déformation spécifiquement à chaque locuteur, comme dans (Serizel & Giuliani, 2014b). De futurs travaux porteront sur la pertinence des paramètres MFCCs pour la parole d'enfant, avec l'étude d'une échelle Mel adaptée aux fréquences des enfants, et la recherche de paramètres spécifiques à nos données d'enfants. Enfin, une modélisation linguistique appropriée pourrait combler l'écart de performance entre les mots correctement lus et ceux contenant des erreurs de fluence ou de déchiffrage.

Références

BOLAÑOS D., COLE R., WARD W., BORTS E. & SVIRSKY E. (2011). FLORA : Fluent oral reading assessment of children's speech. *ACM Trans. Speech Lang. Process.*, **7**(4), 16.

FRINGI E., LEHMAN J. F. & RUSSELL M. J. (2015). Evidence of phonological processes in automatic recognition of children's speech. In *INTERSPEECH*.

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. p. 23–27.

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.

METALLINOU A. & CHENG J. (2014). Using deep neural networks to improve proficiency assessment for children english language learners. In *INTERSPEECH*.

MOSTOW J. & AIST G. (2001). Evaluating tutors that listen : An overview of project listen.

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.

PEDDINTI V., POVEY D. & KHUDANPUR S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. In *INTERSPEECH*.

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L., GLEMBEK O., GOEL N., HANNEMANN M., MOTLICEK P., QIAN Y., SCHWARZ P., SILOVSKY J., STEMMER G. & VESELY K. (2011). The kaldi speech recognition toolkit. In *ASRU*.

POVEY D., PEDDINTI V., GALVEZ D., GHAHREMANI P., MANOHAR V., NA X., WANG Y. & KHUDANPUR S. (2016). Purely sequence-trained neural networks for ASR based on lattice-free MMI. In *INTERSPEECH*, p. 2751–2755.

PROENÇA J. D. L. (2018). *Automatic Assessment of Reading Ability of Children*. Thèse de doctorat.

SERIZEL R. & GIULIANI D. (2014a). Deep neural network adaptation for children's and adults' speech recognition. In *CLiC-it*.

SERIZEL R. & GIULIANI D. (2014b). Vocal tract length normalisation approaches to DNN-based children's and adults' speech recognition. In *SLT*, p. 135–140.

SHI L., BAO F., WANG Y. & GAO G. (2018). Research on transfer learning for Khalkha Mongolian speech recognition based on TDNN. In *IALP*, p. 85–89.

SHIVAKUMAR P. G. & GEORGIOU P. G. (2018). Transfer learning from adult to children for speech recognition : Evaluation, analysis and recommendations. *ArXiv*.

VESELÝ K., GHOSHAL A., BURGET L. & POVEY D. (2013). Sequence-discriminative training of deep neural networks. In *INTERSPEECH*.

WAIBEL A., HANAZAWA T., HINTON G., SHIKANO K. & LANG K. J. (1989). Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **37**(3), 328–339.

WU F., GARCÍA-PERERA L. P., POVEY D. & KHUDANPUR S. (2019). Advances in Automatic Speech Recognition for Child Speech Using Factored Time Delay Neural Network. In *INTERSPEECH*, p. 1–5.

YONG B. & TING H. N. (2011). Speaker-independent vowel recognition for malay children using time-delay neural network. *IFMBE*, **35**.