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

Adaptation de modèles auto-supervisés pour la reconnaissance de phonèmes dans la parole d'enfant

Lucas Block Medin[1, 2], Lucile Gelin[1, 2], Thomas Pellegrini[2]
(1) Lalilo by Renaissance Learning, 236 rue du faubourg Saint Martin, 75010 Paris, France
(2) IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France
lucas.block@renaissance.com, lucile.gelin@renaissance.com,
thomas.pellegrini@irit.fr

RÉSUMÉ

La reconnaissance de parole d'enfant est un domaine de recherche encore peu développé en raison du manque de données et des difficultés caractéristiques de cette tâche. Après avoir exploré diverses architectures pour la RAP d'enfant dans de précédents travaux, nous nous attaquons dans cet article aux nouveaux modèles auto-supervisés. Nous comparons d'abord plusieurs modèles Wav2vec2, HuBERT et WavLM adaptés superficiellement à la reconnaissance de phonèmes sur parole d'enfant, et poursuivons nos expériences avec le meilleur d'entre eux, un WavLM base+. Il est ensuite adapté plus profondément en dégelant ses blocs transformer lors de l'entraînement sur parole d'enfant, ce qui améliore grandement ses performances et le fait surpasser significativement notre modèle de base, un Transformer+CTC. Enfin, nous étudions en détail les comportements de ces deux modèles en conditions réelles de notre application, et montrons que WavLM base+ est plus robuste à diverses tâches de lecture et niveaux de bruit.

ABSTRACT

Adapting self-supervised learning for phoneme recognition in child speech.

Child speech recognition is still an underdeveloped area of research due to the lack of data and the specific difficulties of this task. Having explored various architectures for child speech recognition in previous work, in this article we tackle new self-supervised models. We first compare several Wav2vec2, HuBERT and WavLM models adapted to phoneme recognition in child speech, and continue our experiments with the best of them, a WavLM base+. We then further adapt it by unfreezing its transformer blocks during fine-tuning on child speech, which greatly improves its performance and makes it significantly outperform our base model, a Transformer+CTC. Finally, we study in detail the behaviour of these two models under the real conditions of our application, and show that WavLM base+ is more robust to various reading tasks and noise levels.

MOTS-CLÉS : reconnaissance automatique de la parole ; parole d'enfant ; modèles auto-supervisés.

KEYWORDS: automatic speech recognition, child speech, self-supervised learning.

1 Introduction

Le langage oral des jeunes enfants (5-8 ans) présente des caractéristiques spécifiques liées au développement de leur appareil vocal et à leur contrôle moteur encore en développement : mécanismes articulatoires instables et variabilité spectrale intra- et inter-locuteur·rices (Lee *et al.*, 1999), fréquences fondamentales et formantiques plus élevées que celles des adultes (Mugitani & Hiroya,

2012), ou encore présence d'erreurs phonologiques (Fringi *et al.*, 2015). Ces différences morphologiques et phonologiques constituent les principales raisons des performances limitées des systèmes de reconnaissance automatique de la parole (RAP) lorsqu'ils sont confrontés aux voix d'enfants.

Les outils numériques d'assistance à la lecture ont un impact pédagogique significatif sur les enfants apprenant à lire, et plusieurs initiatives ont été développées au fil du temps (Mostow & Aist, 2001; Bolaños *et al.*, 2011; Godde *et al.*, 2017). Lalilo [1] propose un assistant de lecture destiné aux enfants de 5 à 8 ans, comprenant un exercice de lecture à voix haute qui offre un retour personnalisé grâce au système de reconnaissance automatique de phonèmes présenté dans cet article.

Des recherches antérieures sur la RAP des enfants ont montré des performances inférieures à celles observées chez les adultes (Potamianos & Narayanan, 2003; Shivakumar & Georgiou, 2020; Yeung & Alwan, 2018). Des systèmes hybrides ont démontré des améliorations en combinant données d'adulte et d'enfant (Serizel & Giuliani, 2014), ou en utilisant des techniques d'apprentissage par transfert (Shivakumar & Georgiou, 2020). Les architectures récentes dites bout-à-bout, ou *end-to-end*, ont récemment été adaptées à la RAP d'enfants, et ont atteint ou surpassé les performances des architectures hybrides(Shivakumar & Narayanan, 2021; Gelin *et al.*, 2022).

Récemment, l'apprentissage auto-supervisé (*Self-Supervised Learning*, SSL) a été introduit dans le domaine de la RAP en raison de son grand potentiel pour améliorer les tâches à faibles ressources en exploitant les connaissances préalables acquises à partir de grandes quantités de données non annotées (Mohamed *et al.*, 2022; N *et al.*, 2021). C'est le cas de la RAP d'enfant, où les données sont rares et leur annotation est complexe et coûteuse. De récentes études ont montré que le potentiel d'apprentissage à partir de données non annotées abondantes est élevé pour de la RAP d'enfants (Jain *et al.*, 2023; Fan *et al.*, 2022).

Nous étudions dans ce papier le comportement des modèles auto-supervisés à l'état de l'art sur nos données spécifiques de parole de jeunes enfants apprenant à lire. Après une rapide comparaison de plusieurs modèles (Wav2vec2 (Baevski *et al.*, 2020), HuBERT (Hsu *et al.*, 2021) et WavLM (Chen *et al.*, 2022)) pour la reconnaissance de phonèmes dans la parole d'enfant, nous adapterons plus en profondeur le meilleur d'entre eux, et le comparerons avec notre système actuel, un Transformer+CTC. Nous analyserons en détail les performances de chacun sur différents types de contenu et dans différents niveaux de bruit.

2 Jeux de données

Pour compenser le manque de données disponibles de parole d'enfant, nous utilisons de la parole d'adulte pour entraîner des modèles sources, que nous adaptons ensuite avec de la parole d'enfant.

2.1 Parole d'adulte

Nous utilisons pour notre modèle de base une version du corpus Common Voice [2] français qui contient environ 150 heures de parole lue. Les modèles auto-supervisés sont préentraînés sur de la parole d'adulte non annotée, provenant des corpus suivants :

- Librispeech : 960 heures, annotées, parole lue, anglais (Panayotov *et al.*, 2015) ;
- Libri-Light : 60 000 heures, non annotées, parole lue, anglais (Kahn *et al.*, 2020) ;
- VoxPopuli : 24 000 heures, non annotées, parole multilingue (23 langues) (Wang *et al.*, 2021) ;
- GigaSpeech : 10 000 heures, non annotées, parole lue/spontanée, anglais (Chen *et al.*, 2021).

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 divers types de contenu. Les données sont transcrites manuellement au niveau du mot et chaque mot est étiqueté « correct » ou « incorrect ». Les mots corrects sont phonétisés automatiquement, tandis que les mots incorrects sont transcrits manuellement au niveau du phonème. L'annotation est faite par deux annotateur·rice·s, et l'enregistrement est écarté en cas de désaccord.

Lors de l'apprentissage de la lecture, les élèves effectuent diverses tâches de lecture, de difficulté croissante et appelant à utiliser différents processus cognitifs. Dans la plateforme Lalilo, nous proposons principalement quatre types de contenu, plus ou moins difficiles : des mots isolés, des phrases courtes, des listes de mots et des listes de pseudo-mots. 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. On calcule le niveau de bruit à l'aide d'un rapport signal à bruit (RSB).

Les ensembles d'entraînement et de validation contiennent respectivement 13 heures et 25 minutes de données. Ayant été conçus avant l'ajout de nouveaux types de contenu, ces ensembles ne contiennent que des mots isolés et des phrases. De plus, ils sont uniquement composés d'énoncés correctement prononcés. La transcription correspond au texte demandé à l'élève, phonétisé automatiquement avec un dictionnaire de prononciation. Les ensembles d'entraînement et de validation ont respectivement des RSB moyens de 21,0 dB et 20,6 dB avec des déviations standards de 13,0 dB et 12,6 dB. L'ensemble de test est composé de 3 heures d'énoncés, avec environ 25% des mots qui contiennent une erreur de lecture. Nous utilisons ici les quatre types de contenu, divisant l'ensemble de test en sous-catégories : mots isolés (M, 51 min), phrases (P, 29 min), listes de mots (LM, 56 min) et listes de pseudo-mots (LPM, 50 min). Les valeurs de RSB du test sont identiques à l'ensemble de validation.

3 Description des systèmes

Cette section présente les différents systèmes que nous étudierons dans ce papier. Nous entraînons nos systèmes à la reconnaissance de phonèmes, et non de mots, afin de pouvoir détecter plus efficacement les erreurs de lecture. Tous nos systèmes sont entraînés avec SpeechBrain (Ravanelli *et al.*, 2021).

3.1 Modèle de base : Transformer+CTC

Proposé par (Vaswani *et al.*, 2017) et adapté à la reconnaissance automatique de la parole (RAP) 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, des modules d'auto-attention multi-tête et d'attention croisée, et des réseaux de neurones à propagation avant tenant compte de la position. Le modèle Transformer+CTC est complété par une fonction CTC (*Connectionnist Temporal Classification*) en sortie de l'encodeur, qui permet d'améliorer les performances grâce à un entraînement multi-objectif (entropie croisée et CTC) et un décodage joint attention/CTC (Watanabe *et al.*, 2017; Karita *et al.*, 2019a).

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·es lecteur·rices dans (Gelin *et al.*, 2021, 2022). Notre modèle suit la même architecture que dans les travaux passés cités ci-dessus, mais une nouvelle version a été entraînée avec SpeechBrain

pour faciliter la comparaison avec les autres modèles. Il contient 14,3 millions de paramètres.

Notre modèle Transformer+CTC est entraîné en deux étapes : un premier modèle source est entraîné sur la parole d'adulte provenant du corpus Common Voice, puis ce modèle est adapté par apprentissage par transfert avec le jeu de parole d'enfant Lalilo. Toutes les couches sont ré-entraînées lors de cette seconde étape, comme le conseillent (Shivakumar & Georgiou, 2020) pour de très jeunes enfants.

3.2 Modèles auto-supervisés pré-entraînés

Depuis l'introduction de Wav2vec (Schneider *et al.*, 2019), les modèles auto-supervisés se sont imposés dans le domaine de la RAP. Grâce à l'utilisation de données non annotées pour extraire des représentations latentes, ces modèles peuvent atteindre des résultats à l'état de l'art avec jusqu'à 100 fois moins de données annotées que d'autres modèles supervisés. Ces résultats sont notamment remarquables dans le contexte de la reconnaissance de parole d'enfants, où les modèles de l'architecture Wav2Vec2 (Baevski *et al.*, 2020) atteignent des performances similaires à celles des modèles supervisés de pointe (Jain *et al.*, 2023). Nous avons sélectionné pour notre étude les modèles auto-supervisés préentraînés pour la RAP les plus répandus : Wav2vec2, HuBERT et WavLM.

3.2.1 Wav2vec2

Wav2vec2 (Baevski *et al.*, 2020) est une architecture auto-supervisée dite *end-to-end*, fondée sur des réseaux de neurones convolutifs et transformer. L'architecture peut se diviser en trois grandes parties : un encodeur, un réseau contextuel transformer, et un module de quantification.

L'encodeur se compose de sept blocs contenant une convolution temporelle, suivis d'une couche de normalisation des activations *(Layer Norm)* et d'une fonction d'activation GELU. Le réseau contextuel suit l'architecture Transformer. Il remplace cependant l'encodage positionnel absolu par une couche de convolution, qui agit comme un encodage positionnel relatif. Cet encodage passe par une fonction GELU, est ensuite concaténé aux sorties de l'encodeur, et le tout subit une normalisation *(Layer Norm)*. Le réseau est composé de 12 de ces blocs, de dimension 768, avec une dimension interne de 3082, et 8 têtes d'attention par bloc. Enfin, le module de quantification récupère également la sortie de l'encodeur et la transforme en un ensemble de représentations discrètes via une « quantification » *(product quantization)*.

Le modèle Wav2Vec2 est pré-entraîné pour une tâche de prédiction masquée : il vise à prédire la représentation audio latente quantifiée correcte en contexte d'une utterance malgré l'application d'un masque sur une partie des trames audio. L'objectif global de l'entraînement est de minimiser les fonctions de perte de contraste *(contrastive loss)* et de perte de diversité *(diversity loss)*.

Nous utilisons un modèle acoustique pré-entraîné Wav2vec2.0 Base [3]. Le modèle est entraîné en utilisant le jeu de données LibriSpeech standard 960h (Panayotov *et al.*, 2015).

3.2.2 HuBERT

Le modèle HuBERT (Hsu *et al.*, 2021) reprend l'architecture de Wav2vec2.0, mais remplace le module de quantification par une quantification *K-Means*.

Ce changement implique trois différences fondamentales :

— La représentation discrète est obtenue par la découverte d'unités cachées (*hidden units*), en attribuant à chaque extrait audio un cluster via un algorithme K-Means ;

3. https://huggingface.co/facebook/wav2vec2-base-960h

— L'extraction des représentations est itérative, utilisant d'abord les résultats d'un MFCC, puis les embeddings des couches intermédiaires du modèle pré-entraîné ;
— Les fonctions de perte de contraste et de perte de diversité sont remplacées par une perte d'entropie croisée, ce qui simplifie l'entraînement.

Nous utilisons dans cette étude un modèle acoustique pré-entraîné HuBERT Base [4], entraîné également sur Librispeech standard 960h.

3.2.3 WavLM

L'architecture WavLM (Chen *et al.*, 2022) reprend celle de HuBERT en introduisant un biais de position relative à porte *(gated relative position bias)* dans les mécanismes d'attention. Au lieu de se fier uniquement aux positions absolues des vecteurs clé et requête, le modèle prend ainsi en compte les positions relatives entre ces vecteurs lors du calcul des scores d'attention.

Le modèle WavLM comporte également des modifications dans la phase de pré-entraînement. La tâche de prédiction masquée est remplacée par une tâche de débruitage et prédiction masquée. Ce procédé, qui cherche à rendre le modèle plus robuste, consiste à simuler des entrées bruitées ou de la parole superposée, puis à prédire des pseudo-étiquettes de l'audio original sur la région masquée.

Nous allons étudier dans ce travail deux modèles WavLM :
— Un modèle WavLM Base [5], entraîné avec les mêmes données que les modèles précédents ;
— Un modèle WavLM Base+ [6], qui possède la même architecture que le WavLM Base, mais est entraîné sur un corpus beaucoup plus vaste composé des données LibriLight, GigaSpeech et VoxPopuli, pour un total d'environ 94 000 heures. Ce corpus étendu permet d'améliorer les performances et la robustesse du modèle WavLM tout en gardant un modèle de taille raisonnable (Chen *et al.*, 2022).

4 Adaptation et évaluation des modèles SSL pour la transcription phonémique de parole d'enfant

Nous avons décidé de nous concentrer sur des modèles pré-entraînés SSL de format *Base* plutôt que *Large*. D'une part, la capacité de calcul nécessaire à l'entraînement et au déploiement des modèles *Base* est bien inférieure de par leur plus petit nombre de paramètres (95M contre 317M). D'autre part, nous pouvons constater que, dans le cas de la parole d'enfant, l'amélioration de performance est faible en contrepartie d'une augmentation significative du nombre de paramètres (Jain *et al.*, 2023). Les modèles pré-entraînés en français étaient peu documentés et très hétérogènes en termes de données utilisées, ce qui rendait la comparaison complexe, et nous a poussé à utiliser des modèles anglais.

Pour adapter les systèmes SSL à notre tâche, nous faisons passer les sorties du réseau Transformer (de taille 768) dans une projection linéaire pour faire de la classification de phonèmes. Cette couche comporte 35 classes représentant les phonèmes français et le phonème « vide ». Le modèle est entraîné de façon supervisée avec les données Lalilo, avec pour objectif de minimiser la fonction de perte CTC (*Connectionist Temporal Classification*). Le taux d'erreur phonème (*Phoneme Error Rate*, PER) est utilisé pour mesurer la performance de nos systèmes sur cette tâche.

Dans cette section, nous comparons les différents modèles présentés précédemment et adaptés à la

4. https://huggingface.co/facebook/hubert-base-960h
5. https://huggingface.co/facebook/wavlm-base
6. https://huggingface.co/facebook/wavlm-base-plus

transcription phonétique de parole d'enfant. Nous explorons également deux méthodes d'adaptation en gelant une ou plusieurs parties des modèles, puis étudierons en détail les performances des systèmes en fonction des caractéristiques spécifiques de notre tâche.

4.1 Comparaison des modèles auto-supervisés

Notre première expérience vise à comparer les différents modèles SSL pour notre application. Nous adaptons les modèles en réentraînant uniquement la couche CTC de classification de phonèmes. Les modèles sont entraînés sur maximum 30 epochs, et la sauvegarde obtenant le meilleur PER sur l'ensemble de validation Lalilo est conservée. Pour cette expérience préliminaire, le décodage utilisé est une recherche gloutonne (*greedy search*).

Modèle	**PER**
Wav2vec 2.0	62,9
HuBERT	46,3
WavLM base	46,8
WavLM base+	**41,5**

TABLE 1 – PER (%) des différents modèles entraînés avec *fine-tuning* (avec le corpus Lalilo) de la dernière couche CTC et décodés par *greedy search*

On observe que les performances des modèles HuBERT et WavLM surpassent largement celle du modèle Wav2vec. A quantité de données égale, la différence entre HuBERT et WavLM base n'est pas significative. En revanche, le PER obtenu par le modèle WavLM base+, qui a le même nombre de paramètres mais est entraîné sur 100 fois plus de données, est significativement meilleur. Le reste de l'étude sera en conséquence concentré sur celui-ci.

4.2 Adaptation du modèle WavLM base+

Nous souhaitons aller plus loin dans l'adaptation du modèle WavLM pour améliorer ses performances sur notre application. Plutôt que d'entraîner uniquement la couche CTC avec la parole d'enfant, nous adaptons également une partie du modèle pré-entraîné. Nous suivons pour cela ce qui est fait dans (Jain *et al.*, 2023) pour l'adaptation d'un modèle Wav2vec2 à de la parole d'enfant : pendant les 1000 premières itérations, seule la dernière couche de classification CTC est entraînée, puis le bloc Transformer est également entraîné. L'encodeur CNN, en revanche, reste gelé. Le taux d'apprentissage est fixé à 5e-4 et la taille de batch à 128, suivant les recommandations de (Chen *et al.*, 2022).

Le tableau 2 affiche les valeurs PER obtenues par le modèle de base Transformer+CTC, ainsi que par deux modèles WavLM : celui de la section 4.1 où seule la couche CTC a été adaptée, dit "gelé", et celui adapté plus en profondeur, dit "dégelé". Ici et dans le reste de l'article, le décodage utilisé pour tous les modèles est une recherche par faisceaux (*beam search*) avec une taille de faisceaux de 10.

Modèle	**# params entraînables (# total)**	**PER**
Transformer+CTC	14 M (14 M)	40,5
WavLM base+ "gelé"	28 k (95 M)	39,2
WavLM base+ "dégelé"	90 M (95 M)	**26,1**

TABLE 2 – PER (%) des modèles Transformer+CTC et WavLM base+, décodés par *beam search*

On observe tout d'abord que le modèle WavLM base+ gelé atteint une performance légèrement meilleure que le modèle de base Transformer+CTC, alors que seule sa couche de classification de phonème (moins de 1‰ des poids du modèle) a été entraînée avec la parole d'enfant. Cela montre que les représentations auto-supervisées du modèle adulte WavLM base+, bien que n'ayant pas vu de parole d'enfant lors de leur apprentissage, sont génériques et aisément adaptables à différents types de parole. Ces résultats doivent cependant être nuancés : les deux modèles obtiennent des résultats comparables mais le Transformer+CTC contient près de 7 fois moins de paramètres. En dégelant le bloc Transformer de WavLM base+, on obtient un PER de 26,1%, soit une réduction relative de 33,4% par rapport au modèle gelé. Ce résultat montre que les représentations WavLM peuvent néanmoins être adaptées pour mieux correspondre à une parole spécifique, et que cette adaptation est efficace malgré une petite quantité de données d'adaptation (13 heures).

4.3 Discussion

Dans la section précédente, nous avons vu que les modèles Transformer+CTC et WavLM base+ affichent une différence de PER de 14.4%. Dans cette section, nous souhaitons savoir si cette différence est répartie de façon égale en fonction des différentes tâches de lecture présentées aux systèmes, ainsi qu'en fonction des différentes conditions de bruit en salle de classe.

4.3.1 Application aux tâches de lecture de Lalilo

Nous proposons maintenant d'explorer la performance des systèmes en fonction des différentes tâches de lecture proposées aux élèves, détaillées dans la section 2.2. Les valeurs de PER obtenues sont visibles dans la première partie du tableau 3. On observe aisément que la différence de PER entre les deux modèles dépend effectivement de la tâche de lecture.

Modèle	**Tâche de lecture**				**Niveau de bruit**		
	P	M	LM	LPM	faible	moyen	fort
Transformer+CTC	16,5	34,0	46,5	59,0	14,6	24,6	40,6
WavLM base+ "dégelé"	16,4	25,5	28,3	32,9	13,4	21,7	31,6

TABLE 3 – PER (%) des modèles Transformer+CTC et WavLM base+, décodés par *beam search*, en fonction de la tâche de lecture (P = phrase, M = mot, LM = liste de mots, LPM = liste de pseudo-mots) et/ou du niveau de bruit.

Les phrases courtes représentent la tâche la plus facile pour la RAP, avec un contexte suffisant mais pas trop grand, et la présence de mots de liaisons couramment vus en apprentissage. C'est de plus une tâche classique pour les corpus de parole d'adulte. Sur ce sous-ensemble (P dans le tableau), il n'y a pas de différence significative entre les deux modèles. Les deux modèles sont adaptés avec la même quantité de parole d'enfant. Sur cette tâche facile et connue, l'apprentissage supervisé d'un petit modèle (14 M de paramètres) avec 150 heures de parole d'adulte est donc aussi efficace que l'apprentissage non supervisé d'un gros modèle (95 M) sur près de 100 000 heures de parole.

Sur la reconnaissance de phonèmes dans des mots isolés (M dans le tableau), on observe que le WavLM est significativement meilleur (-8,5% absolu). Les mots pouvant contenir aussi peu que 2 phonèmes, la difficulté du Transformer+CTC peut s'expliquer par le manque de contexte sur lequel les mécanismes d'attention peuvent s'appuyer. Ce phénomène a notamment été observé dans (Chan *et al.*, 2016), où la performance du modèle se dégrade significativement lorsque l'énoncé ne contient qu'un seul mot. Le WavLM contient également un bloc Transformer qui est affecté par ce phénomène, mais il est probablement compensé par l'utilisation d'un encodeur CNN, dont les convolutions permettent de tirer le meilleur parti du manque de contexte.

Les listes de mots (LM) et de pseudo-mots (LPM) n'ont pas été vues pendant l'apprentissage, ce qui en fait des tâches légèrement hors domaine. C'est d'autant plus le cas pour les listes de pseudo-mots car les pseudo-mots n'existent pas et n'ont jamais été vus dans aucun corpus de parole d'adulte ou d'enfant. Sur ces tâches, on observe que le modèle WavLM est bien meilleur que le Transformer+CTC. On observe également que plus la tâche est hors domaine, plus la réduction relative de PER apportée par le WavLM s'accentue : -39% sur les listes de mots, -44% sur les listes de pseudo-mots. On peut déduire de ces observations que le modèle WavLM possède une meilleure capacité de généralisation, qui est sûrement liée à la quantité de données rencontrées, mais également à l'apprentissage auto-supervisé qui est moins contraint et crée ainsi des représentations plus génériques.

4.3.2 Robustesse au bruit de salle de classe

Nous souhaitons également étudier le comportement de nos deux systèmes en conditions réelles de salle de classe, c'est à dire avec différents niveaux de bruit. Nous divisons notre ensemble de test en trois niveaux de bruit :

— Faible : enregistrements avec un RSB supérieur à 25 dB ;
— Moyen : enregistrements avec un RSB compris entre 10 et 25 dB ;
— Fort : enregistrements avec un RSB inférieur à 10 dB.

La seconde partie du tableau 3 affiche les résultats sur ces sous-catégories. On observe évidemment que, pour les deux modèles, la performance se dégrade fortement avec l'augmentation du niveau de bruit. Il est intéressant de noter que la différence de PER entre les deux modèles augmente avec le niveau de bruit : 1,2% sur du bruit faible, 2.9% sur du bruit moyen et 9.0% sur du bruit fort. Le modèle WavLM témoigne ainsi d'une plus grande robustesse au bruit.

Pour confirmer cette observation, nous regardons les performances des modèles en fonction du bruit sur le sous-ensemble de test P, sur lequel les modèles obtiennent un PER comparable.

— Transformer+CTC : 10,6 (faible) - 17,1 (moyen) - 30,0 (fort)
— WavLM base+ : 12,5 (faible) - 17,1 (moyen) - 26,8 (fort)

On voit ainsi que le WavLM est bien plus robuste dans des conditions de bruit fort, au prix d'une moins bonne performance lorsqu'il n'y a qu'un bruit faible. Ces résultats sont en accord avec les changements introduits dans le pré-entraînement de WavLM, ayant pour objectif de rendre le modèle plus robuste à des conditions acoustiques difficiles (Chen *et al.*, 2022).

5 Conclusion

Les systèmes capables de transcrire avec précision la parole d'enfant sont encore rares, notamment en français, en raison d'un manque de données disponibles et d'une difficulté accrue sur ce type de parole. Nous explorons ici l'adaptation des nouveaux modèles auto-supervisés à la reconnaissance de phonèmes sur de la parole de jeunes enfants. Dans un premier temps, nous sélectionnons trois modèles (Wav2vec2, HuBERT et WavLM, en version *base*) et adaptons une couche CTC de classification de phonèmes avec notre corpus de parole d'enfant. Nous observons que les modèles HuBERT et WavLM surpassent Wav2vec2, et que le modèle WavLM base+, entraîné sur 100 fois plus de données tout en conservant le même nombre de paramètres, est significativement plus performant que les autres. Dans un second temps, nous adaptons le modèle WavLM base+ plus en profondeur en dégelant les blocs Transformer du modèle, ce qui améliore ses performances de 33,4% relatifs. Nous montrons qu'il obtient une précision largement meilleure que celle obtenue par notre modèle de base, un Transformer+CTC. Enfin, nous analysons les comportements de nos modèles face à différentes tâches de lecture et conditions de bruit, et montrons que le WavLM base+ est plus efficace sur des enregistrements très courts, généralise mieux à des contenus non vus en apprentissage, et est plus robuste à un bruit de salle de classe.

Références

BAEVSKI A., ZHOU H., MOHAMED A. & AULI M. (2020). wav2vec 2.0 : A framework for self-supervised learning of speech representations.

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. DOI : 10.1145/1998384.1998390.

CHAN W., JAITLY N., LE Q. & VINYALS O. (2016). Listen, Attend and Spell : A neural network for large vocabulary conversational speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 4960–4964. DOI : 10.1109/ICASSP.2016.7472621.

CHEN G., CHAI S., WANG G., DU J., ZHANG W.-Q., WENG C., SU D., POVEY D., TRMAL J., ZHANG J., JIN M., KHUDANPUR S., WATANABE S., ZHAO S., ZOU W., LI X., YAO X., WANG Y., WANG Y., YOU Z. & YAN Z. (2021). Gigaspeech : An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio.

CHEN S., WANG C., CHEN Z., WU Y., LIU S., CHEN Z., LI J., KANDA N., YOSHIOKA T., XIAO X., WU J., ZHOU L., REN S., QIAN Y., QIAN Y., ZENG M., YU X. & WEI F. (2022). Wavlm : Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, **16**, 1–14. DOI : 10.1109/JSTSP.2022.3188113.

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. DOI : 10.1109/ICASSP.2018.8462506.

FAN R., ZHU Y., WANG J. & ALWAN A. (2022). Towards better domain adaptation for self-supervised models : A case study of child asr. *IEEE Journal of Selected Topics in Signal Processing*, **16**(6), 1242–1252. DOI : 10.1109/JSTSP.2022.3200910.

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. (2021). End-to-end acoustic modelling for phone recognition of young readers. *Speech Communication*, **134**, 71–84. DOI : https ://doi.org/10.1016/j.specom.2021.08.003.

GELIN L., PELLEGRINI T., PINQUIER J. & DANIEL M. (2022). Améliorations d'un système Transformer de reconnaissance de phonèmes appliqué à la parole d'enfants apprenants lecteurs. In *34èmes Journées d'Études sur la Parole - Parole, Geste, Musique : des unités à leur organisation (JEP 2022)*, volume Session Posters n° 2, p. à paraître, Noirmoutier, France : Association Francophone de la Communication Parlée. HAL : hal-03898401.

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. DOI : 10.21437/WOCCI.2017-4.

HSU W.-N., BOLTE B., TSAI Y.-H. H., LAKHOTIA K., SALAKHUTDINOV R. & MOHAMED A. (2021). Hubert : Self-supervised speech representation learning by masked prediction of hidden units.

JAIN R., BARCOVSCHI A., YIWERE M. Y., BIGIOI D., CORCORAN P. & CUCU H. (2023). A wav2vec2-based experimental study on self-supervised learning methods to improve child speech recognition. *IEEE Access*, **11**, 46938–46948. DOI : 10.1109/ACCESS.2023.3275106.

KAHN J., RIVIERE M., ZHENG W., KHARITONOV E., XU Q., MAZARE P., KARADAYI J., LIPTCHINSKY V., COLLOBERT R., FUEGEN C., LIKHOMANENKO T., SYNNAEVE G., JOULIN A., MOHAMED A. & DUPOUX E. (2020). Libri-light : A benchmark for asr with limited or no supervision. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* : IEEE. DOI : 10.1109/icassp40776.2020.9052942.

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. DOI : 10.21437/Interspeech.2019-1938.

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. DOI : 10.1109/ASRU46091.2019.9003750.

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. DOI : 10.1121/1.426686.

MOHAMED A., LEE H.-Y., BORGHOLT L., HAVTORN J. D., EDIN J., IGEL C., KIRCHHOFF K., LI S.-W., LIVESCU K., MAALØE L. *et al.* (2022). Self-supervised speech representation learning : A review. *IEEE Journal of Selected Topics in Signal Processing*.

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. DOI : 10.5555/570950.570957.

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. DOI : 10.1250/ast.33.215.

N K. D., WANG P. & BOZZA B. (2021). Using Large Self-Supervised Models for Low-Resource Speech Recognition. In *Proc. Interspeech 2021*, p. 2436–2440. DOI : 10.21437/Interspeech.2021-631.

PANAYOTOV V., CHEN G., POVEY D. & KHUDANPUR S. (2015). Librispeech : An asr corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 5206–5210. DOI : 10.1109/ICASSP.2015.7178964.

POTAMIANOS A. & NARAYANAN S. (2003). Robust Recognition of Children's Speech. *IEEE Transactions on Speech and Audio Processing*, **11**(November 2003), 603–616. DOI : 10.1109/TSA.2003.818026.

RAVANELLI M., PARCOLLET T., PLANTINGA P., ROUHE A., CORNELL S., LUGOSCH L., SUBAKAN C., DAWALATABAD N., HEBA A., ZHONG J., CHOU J.-C., YEH S.-L., FU S.-W., LIAO C.-F., RASTORGUEVA E., GRONDIN F., ARIS W., NA H., GAO Y., MORI R. D. & BENGIO Y. (2021). SpeechBrain : A general-purpose speech toolkit. arXiv :2106.04624.

SCHNEIDER S., BAEVSKI A., COLLOBERT R. & AULI M. (2019). wav2vec : Unsupervised Pre-Training for Speech Recognition. In *Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Graz*, p. 3465–3469. DOI : 10.21437/Interspeech.2019-1873.

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. HAL : hal-01393975.

SHIVAKUMAR P. G. & GEORGIOU P. (2020). Transfer learning from adult to children for speech recognition : Evaluation, analysis and recommendations. *Computer Speech & Language*, **63**, 101077. DOI : 10.1016/j.csl.2020.101077.

SHIVAKUMAR P. G. & NARAYANAN S. (2021). End-to-end neural systems for automatic children speech recognition : An empirical study. *ArXiv preprint :2102.09918*.

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.

WANG C., RIVIÈRE M., LEE A., WU A., TALNIKAR C., HAZIZA D., WILLIAMSON M., PINO J. & DUPOUX E. (2021). Voxpopuli : A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation.

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. DOI : 10.1109/JSTSP.2017.2763455.

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. DOI : 10.21437/Interspeech.2018-2297.