Interspeech 2024
1-5 September 2024, Kos, Greece

Self-Supervised Models for Phoneme Recognition: Applications in Children's Speech for Reading Learning

Lucas Block Medin[1,2], *Thomas Pellegrini*[2], *Lucile Gelin*[1,2]

[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, thomas.pellegrini@irit.fr, lucile.gelin@renaissance.com

Abstract

Child speech recognition is still an underdeveloped area of research due to the lack of data (especially on non-English languages) and the specific difficulties of this task. Having explored various architectures for child speech recognition in previous work, in this article we tackle recent self-supervised models. We first compare wav2vec 2.0, HuBERT and WavLM models adapted to phoneme recognition in French child speech, and continue our experiments with the best of them, 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.
Index Terms: speech recognition, child speech, self-supervised learning

1. Introduction

Reading tutors have a significant pedagogical impact on children learning to read, and several initiatives have been developed over time [1, 2, 3]. We have created a reading assistant for children aged 5 to 8, including a read-aloud exercise that provides personalised feedback thanks to the automatic phoneme recognition system presented in this article.

The oral language of young children (5-8 years old) has specific characteristics linked to the development of their vocal apparatus and motor control capacities: unstable articulatory mechanisms and intra- and inter-speaker spectral variability [4], fundamental and higher formant frequencies [5], and the presence of phonological errors [6]. These morphological and phonological differences, as well as the lack of child speech data, are the main reasons for the limited performance of automatic speech recognition (ASR) systems for children [7, 8, 9].

Hybrid deep neural network - Hidden Markov Model systems obtained improvements by combining adult and child data [10], or by using transfer learning techniques [8]. Supervised end-to-end architectures have recently been adapted to child ASR, and have reached or surpassed the performance of hybrid architectures [11, 12]. Recently, self-supervised learning (SSL) has been introduced into the field of ASR because of its great potential to improve low-resource tasks by exploiting prior knowledge acquired from large amounts of unlabeled data [13]. This is the case in child ASR, where data is scarce and annotation is complex and expensive. Recent studies have shown that the learning potential from abundant unlabeled data is high for child ASR [14, 15, 16].

Our educational use case however demonstrates specific difficulties, such as young age (5-8), unusual tasks (pseudoword reading), non-English language and classroom noise. The impact of SSL models has not yet been studied for child speech with those specificities. Our contributions on this paper are therefore multiple:

- We compare the performance of state-of-the-art SSL models (wav2vec 2.0, HuBERT, WavLM), and show the adaptability of these adult-trained English model for phoneme recognition on young children's speech in another language (French) ;
- We explore several finetuning processes for adapting the new WavLM model to our application with a small quantity of French children's speech data, and significantly outperform our baseline model, a supervised Transformer+CTC;
- We demonstrate that the WavLM base+ model displays better generalization capabilities and better robustness to noise than our supervised model.

2. Speech material

2.1. Adult speech

For our baseline model, we use a version of the French Common Voice[1] which contains around 150 hours of read speech. The self-supervised models are trained and finetuned on adult speech from the following corpora:

- Librispeech [17]: 960 hours transcribed English read speech;
- LibriLight [18]: 60k hours unlabeled English read speech;
- VoxPopuli [19]: 24k hours unlabeled English speech;
- GigaSpeech [20]: 10k hours unlabeled English read and spontaneous speech.

2.2. Child speech

2.2.1. In-house child speech corpus in French

Our in-house (IH) corpus contains recordings of French children from 1st to 3rd grade (age 5- 8), reading aloud various types of content. The data is transcribed manually at word level and each word is labelled "correct" or "incorrect". Correct words are automatically phonetized, while incorrect words are manually transcribed at the phoneme level. Annotation is done by two annotators, and the recording is discarded when they disagree.

When learning to read, pupils perform various reading tasks, requiring them to use different cognitive processes. In our reading platform, we offer four main types of content, with varying degrees of difficulty: isolated words, short sentences, word lists and pseudoword lists. The recordings are mainly collected as part of the oral reading exercise on the platform,

[1]Corpus available on : https://voice.mozilla.org/fr

10.21437/Interspeech.2024-1095

which is most often used in classrooms under reduced supervision: they contain varying levels of babble noise. The noise level is calculated using a signal-to-noise ratio (SNR).

The training and validation sets respectively contain 13 hours and 25 minutes of data. Having been designed before the addition of new types of content, these sets contain only isolated words and sentences. In addition, they are composed solely of correctly read utterances. The transcription corresponds to the text requested from the student, automatically phonetized using a pronunciation dictionary. The training and validation sets have mean SNRs of 21.0 ± 13.0 dB and 20.6 ± 12.6 dB respectively. The test set consists of 3 hours of utterances, with approximately 25% of the words containing a reading error. We use all four content types here, dividing the test set into subcategories: isolated words (W, 51 min), sentences (S, 29 min), word lists (WL, 56 min) and pseudoword lists (PWL, 50 min). The test's SNR values are identical to the validation set.

2.2.2. MyST

The My Science Tutor (MyST) Children's Speech Corpus [21] is a large-scale collection of spontaneous American English conversational speech between 3rd, 4th, and 5th grade students and a virtual science tutor. Approximately 45% of all utterances, amounting to 224 hours of speech, have been transcribed at the word level and are presented in .trn file format.

To prepare the MyST corpus for phonetization, several data cleaning steps were performed. First, all utterances containing typographical transcription errors or non-word labels such as DISCARD, SILENCE, and NO_SIGNAL were removed. Secondly, filled pauses, non-speech events, truncated words, and unintelligible words were deleted from the transcriptions. After cleaning, we ended up with a train set of 161 hours, a development set of 25 hours, and a test set of 27 hours. We used the splits provided by MyST for these datasets.

3. Systems description

This section presents the different systems we will be studying in this paper. We train our systems to recognise phonemes, not words, so that we can detect reading errors more effectively. All our systems are trained with SpeechBrain [22].

3.1. Baseline system: Transformer+CTC with F-bank

Proposed by [23] and adapted to automatic speech recognition (ASR) by [24], the Transformer model follows a sequence-to-sequence encoder-decoder end-to-end architecture. It is based solely on attention mechanisms and compensates the lack of recurrence with positional encodings, multi-head self-attention and cross-attention modules, and position-wise feed-forward neural networks. The Transformer+CTC model is complemented by a CTC (*Connectionnist Temporal Classification*) function at the encoder output, which improves performance through multi-objective training (cross-entropy and CTC) and joint attention/CTC decoding [25, 26].

The choice of this architecture is based on its excellent performance in adult speech recognition tasks [27], which was confirmed on the speech of children learning to read in our previous work [28, 12]. Our model uses Mel F-bank features as input, and follows the architecture used in the papers cited above. Our Transformer+CTC model is trained in two stages: firstly trained on adult speech from the Common Voice corpus, then adapted with transfer learning with the IH child speech corpus. All the layers are re-trained during this second stage, as advised by [8] for very young children.

3.2. Self-supervised models

By using unlabeled data to extract latent representations, self-supervised models can achieve state-of-the-art results with up to 100 times less annotated data than other supervised models. These results are particularly noteworthy in the context of children's speech recognition, where models based on the wav2vec 2.0 [29] architecture achieve performance similar to that of state-of-the-art supervised models [14]. For our study, we selected the most widely used self-supervised pre-trained models for ASR: wav2vec 2.0, HuBERT and WavLM.

3.2.1. wav2vec 2.0

wav2vec 2.0 [29] is a self-supervised end-to-end architecture based on convolutional and transformer neural networks. The architecture can be divided into three main parts: an encoder, a contextual transformer network, and a quantization module.

The encoder consists of seven temporal convolution blocks, followed by an activation normalisation layer and a GELU activation function. It replaces the absolute positional encoding with a convolution layer, which acts as a relative positional encoding. This encoding is fed through a GELU function, then concatenated with the encoder outputs, and the whole undergoes a layer normalisation. Finally, the quantization module also takes the encoder output and transforms it into a set of discrete representations via product quantization.

The wav2vec 2.0 model is pre-trained for a masked prediction task: it aims at predicting the correct quantized latent audio representation in context of an utterance despite the application of a mask on part of the audio frames. The overall training objective is to minimise the contrast and diversity loss functions.

We use a wav2vec 2.0 Base model[2]. The model is trained using the standard LibriSpeech 960h dataset [17].

3.2.2. HuBERT

The HuBERT [30] model uses the same architecture as wav2vec 2.0, but replaces the quantization module with a K-Means quantization, which involves three fundamental changes:

- Discrete representations of speech segments are obtained by discovering hidden units, by assigning a cluster to each audio extract using a K-Means algorithm;
- The extraction of representations is iterative, using first the results of an MFCC, then the embeddings of the intermediate layers of the pre-trained model;
- The contrast loss and diversity loss functions are replaced by a cross-entropy loss, which simplifies training.

In this study, we use a pre-trained HuBERT Base acoustic model[3], also trained on the standard Librispeech 960h.

3.2.3. WavLM

The WavLM [31] architecture follows HuBERT's, while introducing a gated relative position bias into the attention mechanisms. Instead of relying solely on the absolute positions of the key and query vectors, the model considers the relative positions between these vectors when calculating attention scores.

The WavLM model also includes modifications in the pre-training phase. The masked prediction task is replaced by a

[2] https://huggingface.co/facebook/wav2vec2-base-960h
[3] https://huggingface.co/facebook/hubert-base-ls960

masked denoising and prediction task. This process, which aims at making the model more robust, involves simulating noisy inputs or overlapping speech, then predicting pseudo-labels of the original audio over the masked region. We will study two WavLM models:

- A WavLM Base model[4], trained with the same data as the previous models;
- A WavLM Base+ model[5], which has the same architecture as WavLM Base, but is trained on a much larger corpus consisting of LibriLight, GigaSpeech and VoxPopuli data, for a total of around 94,000 hours. This extended corpus makes it possible to improve the performance and robustness of the WavLM model while maintaining a reasonable model size [31].

4. Model adaptation and evaluation for child speech phoneme transcription

We decided to focus on pre-trained SSL models in the *Base* format rather than *Large*. On the one hand, the computing capacity required to train and deploy the *Base* models is much lower due to their smaller number of parameters (95M versus 317M). On the other hand, we can see that, in the case of child speech, the performance improvement is small at the cost of a significant increase in the number of parameters [14]. The French pre-trained models were poorly documented and very heterogeneous in terms of the data used, which made the comparison complex, and led us to use English models.

To adapt the SSL systems to our task, we feed the output of the Transformer network (size 768) in a linear projection for phoneme classification. This layer is composed of 35 classes representing the French phonemes and the "empty" phoneme. The model is trained in a supervised manner with the IH corpus, with the aim of minimising the CTC loss function (*Connectionist Temporal Classification*). Phoneme error rate (PER) is used to measure the performance of our systems on this task.

4.1. Comparing SSL models for our application

Our first objective is to compare available SSL models for our application. We adapt the models by re-training only the CTC phoneme classification layer with the IH corpus. The models are trained on 30 epochs, on AWS g4dn.xlarge GPUs for 10 hours, which consumes 325 gCO2eq[6] per training. The checkpoint with the best PER on the IH validation set is retained. For this preliminary experiment, we decode with a greedy search.

Table 1: *PER obtained on IH test set with different models with the CTC layer fine-tuned on the IH corpus (greedy search)*

Model	PER (%)
wav2vec 2.0	62.9
HuBERT	46.3
WavLM base	46.8
WavLM base+	**41.5**

The HuBERT and WavLM models significantly outperform wav2vec 2.0, likely due to their use of K-means clustering for quantization. This approach improves generalization by learning discrete representations that capture high-level acoustic patterns from the unlabeled data. While HuBERT and WavLM base show comparable performance when trained on the same amount of data, the WavLM base+ model achieves a substantially lower PER. This improvement can be attributed to its pre-training on 100 times more data, which allows the K-means clustering to discover more robust and generalizable representations. The increased generalization ability makes WavLM base+ a promising candidate for further exploration in this study.

4.2. Adapting WavLM base+ model with in-house data

We go further in adapting the WavLM model to improve its performance on our application. Rather than just training the CTC layer with child speech, we also adapt part of the pre-trained model. To do this, we follow what is done in [14] for adapting a wav2vec 2.0 model to children's speech: for the first 1000 iterations, only the last CTC classification layer is trained, then the Transformer block is also trained. The CNN encoder, on the other hand, remains frozen. The learning rate is set to 5e-4 and the batch size to 128, following recommendations in [31]. The model is trained on 55 epochs on an AWS g4dn.12xlarge GPUs (8.5 hours, 1.96 kgCO2eq).

Table 2 displays the PER values obtained by the baseline Transformer+CTC model, and two WavLM models: the one in the 4.1 section where only the CTC layer has been adapted and the one adapted at greater depth, respectively denominated as "frozen" and "full". Here and in the following, the decoding used for all the models is a beam search (size 10).

Table 2: *PER obtained on IH test set with Transformer+CTC and WavLM base+ models fine-tuned on IH data (beam search)*

Model	# train. params (# total)	PER (%)
Transformer+CTC	14 M (14 M)	40.5
WavLM base+ IH-frozen	28 k (95 M)	39.2
WavLM base+ IH-full	90 M (95 M)	**26.1**

We first observe that the WavLM base+ IH-frozen model achieves a slightly better performance than the Transformer+CTC model, while only its phoneme classification layer (less than 1‰ of the model weights) has been trained with child speech. This shows that the self-supervised representations of the WavLM base+ adult model, despite not having seen child speech when trained, are generic and easily adaptable to different speech types. Also, WavLM having been trained on English data, we can deduce that the representations well adapt to other languages. However, these results must be taken with precautions: the two models obtain comparable results, but the Transformer+CTC contains almost 7 times less parameters. Unfreezing the Transformer block of WavLM base+ gives a PER of 26.1%, a relative reduction of 33.4% compared with the frozen model. This result shows that WavLM representations can nevertheless be adapted to better match a specific kind of speech, and that this adaptation is effective despite a small amount of adaptation data (13 hours).

4.3. Leveraging other child speech data

Having only a small amount of child speech data, we want to explore leveraging larger child speech datasets to improve our models performance. In this section, we "full" fine-tune our pre-trained adult WavLM base+ model with the MyST train dataset (161 hours). The training lasted 34 hours on AWS g4dn.12xlarge GPUs (7.86 kgCO2eq). We validate this model by testing it on the MyST evaluation corpus, on which it obtains a very good PER of 11.8%. This result compares to the

[4] https://huggingface.co/microsoft/wavlm-base

[5] https://huggingface.co/microsoft/wavlm-base-plus

[6] https://engineering.teads.com/sustainability/carbon-footprint-estimator-for-aws-instances/

one obtained in [16] on the MyST corpus with a smaller test set (recordings shorter than 15 seconds). We then apply frozen and full fine-tuning with the IH dataset.

Table 3: *PER obtained on IH test set with WavLM base+ models fine-tuned on MyST then IH data*

Model	PER (%)
WavLM base+ MyST-IH-frozen	58.8
WavLM base+ MyST-IH-full	36.3

We can see in Table 3 that using the MyST data does not bring the expected improvements. We can form several hypothesis. First, the two datasets might be too different: MyST students are older (which makes their voice significantly different [9]), speech is spontaneous versus read, and the MyST vocabulary is quite specialized. This would explain that using a model fine-tuned on adult read speech works better than using a model fine-tuned on far-from-domain child speech. Secondly, the models have been first pre-trained and fine-tuned on adult English speech, then fine-tuned on child English speech, to be finally tested on child French speech. Its bad performance could be due to an over-fitting on the English phoneme representations. The discussion will thus be held on our best model, the WavLM base+ IH-full, that obtains the lowest PER (26.1%).

5. Discussion

In the previous section, we saw that the Transformer+CTC and our best WavLM base+ model show a PER difference of 14.4%. In this section, we want to find out whether this difference is evenly distributed across diverse reading tasks and noise levels.

5.1. Application to diverse reading tasks

We now explore the performance of the systems according to the different reading tasks proposed to the students (see section 2.2.1). We easily see in table 4 that the PER difference between the two models does indeed depend on the reading task.

Table 4: *PER (%) obtained with Transformer+CTC and WavLM base+ IH-full models, depending on reading tasks (S = sentence, W = word, WL = words list, PWL = pseudowords list)*

Model	Reading task			
	S	W	WL	PWL
Transformer+CTC	16.5	34.0	46.5	59.0
WavLM base+ IH-full	16.4	25.5	28.3	32.9

Short sentences (S in the table) represent the easiest task for ASR, with a sufficient but not too large context, and the presence of linking words commonly seen in training. It is also a classic task for adult speech corpora. Both models perform similarly, indicating that on an easy and well-known task, supervised learning of a small model (14M parameters) with 150 hours of adult speech is as effective as unsupervised learning of a large model (95M) with almost 100,000 hours of speech.

For phoneme recognition in isolated words (W), WavLM is significantly better (-8.5% absolute). Since words can contain as few as 2 phonemes, the Transformer+CTC's difficulty can be explained by the lack of context, hurting the attention mechanisms. This phenomenon was notably observed in [32], where the model's performance degrades significantly when the utterance contains only a single word. WavLM also contains a Transformer block which is affected by this phenomenon, but it is probably compensated for by the use of a CNN encoder, whose convolutions make the most of the lack of context.

The word lists (WL) and pseudoword lists (PWL) were not seen during training, making them slightly out-of-domain. This is even more the case for the pseudoword lists because pseudowords do not exist and have never been seen in any corpus of adult or child speech. WavLM model clearly outperforms the Transformer+CTC on these tasks. We also observe that the more outside the domain is the task, the greater is the relative reduction in PER provided by WavLM: -39% on word lists, -44% on pseudoword lists. We can deduce from these observations that the WavLM model has a better generalisation capacity, which is undoubtedly linked to the quantity of data encountered, but also to self-supervised learning, which is less constrained and thus creates more generic representations.

5.2. Robustness to classroom noise

We also want to study the behaviour of our two systems in real classroom conditions. We divide our test set into three noise levels, and compute the PER on each (table 5): low noise (SNR above 25 dB), medium noise (SNR between 10 and 25 dB) and high noise (SNR below 10 dB).

Table 5: *PER (%) obtained with Transformer+CTC and WavLM base+ IH-full models, depending on the noise level*

Model	Noise level		
	low	medium	high
Transformer+CTC	14.6	24.6	40.6
WavLM base+ IH-full	13.4	21.7	31.6

It is clear that, for both models, performance deteriorates sharply with increasing noise level. Interestingly, the difference in PER between the two models increases with noise level: 1.2% at low noise, 2.9% at medium noise and 9.0% at high noise. The WavLM model is therefore more robust to noise.

To confirm this observation, we look at the performance of the models as a function of noise level on the sentence test subset, on which the models obtain a comparable PER.

- Transformer+CTC : 10.6 (low) - 17.1 (medium) - 30.0 (high)
- WavLM base+ : 12.5 (low) - 17.1 (medium) - 26.8 (high)

This shows that WavLM is more robust in high noise conditions, at the cost of poorer performance in low noise conditions. These results are in line with the changes introduced in the WavLM pre-training, aimed at making the model more robust to difficult acoustic conditions [31].

6. Conclusion

Accurate transcription of children's speech remains challenging, particularly in French, due to data scarcity and the inherent difficulties of this speech type. In this work, we explored adapting self-supervised models to phoneme recognition in young children's speech. When comparing wav2vec 2.0, HuBERT and WavLM *Base* models by training only a CTC phoneme classification layer, we found that HuBERT and WavLM outperformed wav2vec 2.0. The WavLM base+ model, trained on significantly more data, achieved the best performance. Further adapting WavLM base+ by unfreezing the Transformer blocks during fine-tuning improved its accuracy by 33.4%, reaching a 26.1% PER and surpassing our Transformer+CTC baseline. Experiments leveraging additional child speech data to overcome data scarcity yielded negative results, potentially due to domain and language mismatches. Analyzing model performance on various tasks and noise conditions, we demonstrated WavLM base+'s superiority on short recordings, better generalization to unseen content, and increased robustness to classroom noise.

7. References

[1] J. Mostow and G. Aist, "Evaluating tutors that listen: An overview of Project LISTEN," in *Smart machines in education: The coming revolution in educational technology.* The MIT Press, 2001, pp. 169–234.

[2] D. Bolaños, R. Cole, W. Ward, E. Borts, and E. Svirsky, "FLORA: Fluent oral reading assessment of children's speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 7, no. 4, p. 16, 2011.

[3] E. Godde, G. Bailly, D. Escudero, M.-L. Bosse, and G. Estelle, "Evaluation of reading performance of primary school children: Objective measurements vs. subjective ratings," in *Proc. of the International Workshop on Child Computer Interaction (WOCCI)*, 2017, pp. 23–27.

[4] S. Lee, A. Potamianos, and S. S. Y. Narayanan, "Acoustics of children's speech: developmental changes of temporal and spectral parameters," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999.

[5] R. Mugitani and S. Hiroya, "Development of vocal tract and acoustic features in children," *The Journal of the Acoustical Society of Japan*, vol. 68, no. 5, pp. 234–240, 2012.

[6] E. Fringi, J. F. Lehman, and M. J. Russell, "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*, 2015, pp. 1621–1624.

[7] A. Potamianos and S. Narayanan, "Robust Recognition of Children's Speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. November 2003, pp. 603–616, 2003.

[8] P. G. Shivakumar and P. Georgiou, "Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations," *Computer Speech & Language*, vol. 63, p. 101077, 2020.

[9] G. Yeung and A. Alwan, "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*, 2018, pp. 1661–1665.

[10] R. Serizel and D. Giuliani, "Deep neural network adaptation for children's and adults' speech recognition," in *Proc. of the Italian Computational Linguistics Conference (CLiC-it)*, 2014, pp. 137–140.

[11] P. G. Shivakumar and S. Narayanan, "End-to-end neural systems for automatic children speech recognition: An empirical study," *ArXiv preprint:2102.09918*, 2021.

[12] L. Gelin, T. Pellegrini, J. Pinquier, and M. Daniel, "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 (JEP 2022)*. France: Association Francophone de la Communication Parlée, Jun. 2022.

[13] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe *et al.*, "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[14] R. Jain, A. Barcovschi, M. Y. Yiwere, D. Bigioi, P. Corcoran, and H. Cucu, "A wav2vec2-based experimental study on self-supervised learning methods to improve child speech recognition," *IEEE Access*, vol. 11, pp. 46 938–46 948, 2023.

[15] R. Fan, Y. Zhu, J. Wang, and A. Alwan, "Towards better domain adaptation for self-supervised models: A case study of child asr," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1242–1252, 2022.

[16] J. Li, M. Hasegawa-Johnson, and N. L. McElwain, "Analysis of self-supervised speech models on children's speech and infant vocalizations," *arXiv preprint arXiv:2402.06888*, 2024.

[17] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[18] J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P. Mazare, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux, "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, May 2020.

[19] C. Wang, M. Rivière, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," 2021.

[20] G. Chen, S. Chai, G. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang, M. Jin, S. Khudanpur, S. Watanabe, S. Zhao, W. Zou, X. Li, X. Yao, Y. Wang, Y. Wang, Z. You, and Z. Yan, "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," 2021.

[21] S. S. Pradhan, R. A. Cole, and W. H. Ward, "My science tutor (myst) – a large corpus of children's conversational speech," 2023.

[22] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. of the International Conference on Neural Information Processing Systems (NIPS)*. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.

[24] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5884–5888.

[25] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/Attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[26] S. Karita, N. E. Y. Soplin, S. Watanabe, M. Delcroix, A. Ogawa, and T. Nakatani, "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*, 2019, pp. 1408–1412.

[27] S. Karita, X. Wang, S. Watanabe, T. Yoshimura, W. Zhang, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, and R. Yamamoto, "A Comparative Study on Transformer vs RNN in Speech Applications," *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, no. April 2020, pp. 449–456, 2019.

[28] L. Gelin, M. Daniel, J. Pinquier, and T. Pellegrini, "End-to-end acoustic modelling for phone recognition of young readers," *Speech Communication*, vol. 134, pp. 71–84, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639321000959

[29] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.

[30] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," 2021.

[31] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, M. Zeng, X. Yu, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, pp. 1–14, 10 2022.

[32] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, Attend and Spell: A neural network for large vocabulary conversational speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4960–4964.