In recent years, the remarkable advancements in Automatic Speech Recognition (ASR) technology, have sparked the emergence of various applications across diverse domains. Among the potential beneficiaries, children stand out as a significant demographic. ASR technology offers promise for innovative interactions between humans and machines through voice, automated reading tutors, and speech pathologist assistants, among other applications. However, the unique challenges presented by children's speech impede the performance of existing ASR systems designed for adults. Consequently, there exists an urgent need for tailored solutions that address the specific challenges posed by children's speech. This necessitates innovative approaches and methodologies to adapt ASR systems effectively to the nuances of children's speech. The research presented in this thesis has made significant strides in this direction, contributing valuable insights and methodologies toward the development of ASR systems optimized for children, thus advancing the field and unlocking new possibilities for using ASR technology to benefit this important demographic.

In the initial phase of this thesis, we lay the foundation for our research by providing an overview of the context and the formidable challenges associated with recognising children's speech. The primary challenge is characterised by the significant variability present in both the acoustic and linguistic components of children’s speech, compounded by intra- and inter-speaker variability. Notably, children's speech encompasses frequency ranges that are shifted and broadened compared to those found in adult speech. Furthermore, the intricate process of language acquisition in young children introduces an additional layer of complexity, making accurate recognition challenging for both human listeners and automated systems. Addressing these multifaceted challenges inherent in children’s speech recognition requires access to a substantial volume of data, thereby constituting a secondary major challenge. While datasets for adult speech are increasingly abundant in size, children’s speech corpora remain scarce and typically of limited size. In this preliminary part of the thesis, we delve into various methodologies employed in the literature to address the diverse challenges associated with children's ASR. Through an extensive literature review, we identify the most promising approaches, which serve as the basis for further development in this thesis. Additionally, we provide a non-exhaustive compilation of children's speech corpora documented in the literature, representing the most comprehensive collection available to date. This compilation serves as a valuable resource for researchers, offering insight into existing datasets for further investigation and experimentation in the context of children’s speech recognition.

In the second segment of the thesis, our focus shifted towards the development and exploration of a hybrid hidden Markov model-based ASR system specifically tailored for children's speech recognition. This emphasis extended to both the English language and a low-resource language, particularly European Portuguese. Our endeavours revolved around the comprehensive evaluation of various knowledge transfer approaches, with a particular emphasis on multi-task and transfer learning, to assess their efficacy in the realm of children's ASR. Among the diverse approaches evaluated, transfer learning emerged as the most potent technique for systems dedicated solely to recognising children's speech. Multi-task learning, on the other hand, demonstrated effectiveness in scenarios where the system needed to concurrently recognise both children's and adult's speech. Additionally, we introduced a novel approach, called “multilingual transfer learning”, which combines elements of both multi-task and transfer learning methodologies. Our findings underscored the efficacy of training a multilingual children's ASR system as a superior initialisation for subsequent transfer learning on a target children's dataset, particularly in low-resource settings. This approach proved instrumental in mitigating the challenges associated with limited data availability, paving the way for more robust and accurate children's speech recognition systems across diverse linguistic contexts.

In the subsequent phase of this thesis, our focus shifted towards exploring the end-to-end paradigm, aiming to advance the current state-of-the-art approaches in children's ASR. Departing from the conventional approach of transfer learning over the entire model, we proposed a more nuanced evaluation strategy. Our investigation unveiled the pivotal role of the Encoder in the fine-tuning process for end-to-end children's ASR. This finding resonates with the understanding that, within the context of children's speech, acoustic variability significantly outweighs linguistic factors in contributing to the degradation of recognition accuracy. Moreover, our research shed light on the effectiveness of targeting higher layers, situated closer to the output of the Encoder. These insights offer valuable recommendations for optimising the development of children's ASR models through transfer learning. In this section of the thesis, we also introduced the novel concept of "partial fine-tuning" for Transformer-based architectures. Our findings indicated that fine-tuning specific components of the network outperformed the traditional approach of fine-tuning the entire model. Notably, the Feed-Forward Network component emerged as the most crucial module, yielding a remarkable score. This innovative approach holds promise for enhancing the performance of children's ASR systems by focusing fine-tuning efforts on the most impactful network components.

Next, motivated by the need for parameter-efficient knowledge transfer, particularly in scenarios with limited training data, we delved into the use of Adapter modules. These modules, comprising two linear layers integrated into a pre-trained frozen model, offer a mechanism for knowledge transfer while retaining the weights and knowledge encapsulated in the pre-trained model. Our investigation encompassed the evaluation of various configurations within both Transformer and Conformer architectures. Among the myriad configurations evaluated, the parallel configuration, along with its Conformer extension known as Two Parallel Adapters, emerged as the best for transferring knowledge to children's speech. Notably, these configurations surpassed the performance of entire model fine-tuning, achieving superior results while using only 10% of the parameters involved in traditional fine-tuning. This evaluation underscored the promise of Adapters in the context of children's ASR, suggesting their potential for more precise adaptation. To further enhance adaptability, we introduced an unsupervised procedure wherein utterances were clustered using k-means applied to their respective speaker embedding. This method is justified by the idea that speech with comparable acoustic characteristics, as detected by the speaker embedding extractor, would be better if an Adapter trained on similar speech patterns is used as opposed to a general children's Adapter.

Expanding on the effectiveness of Adapters in bridging the gap between the source and target domains for children's speech, we leveraged Adapters to enhance data augmentation with imperfect data for children's ASR. Specifically, we introduced the innovative "Double Way Adapter Tuning" method, which uses Text-to-Speech (TTS) data as data augmentation. Recognising that synthetic speech often exhibits acoustic mismatches with real speech, it was imperative to mitigate this gap, a challenge our "Double Way Adapter Tuning" approach aimed to tackle. This method comprised a two-step procedure: initially training Adapter modules using imperfect TTS data, followed by fine-tuning both Adapters and the entire model weights using a combination of synthetic and real data. Notably, the data underwent a distinct dual-pathway approach, with synthetic speech passing through the Adapters while real speech bypassed them. The implementation of the "Double Way Adapter Tuning" approach yielded significant improvements over baseline and previous techniques across both Transformer and Conformer architectures, highlighting the efficacy of our method. Furthermore, we extended the speaker embedding filtering of imperfect synthetic data by incorporating x-vectors instead of i-vectors. This involved using cosine similarity between the reference and generated utterances to discard utterances that may have been incorrectly generated, thereby enhancing the quality and reliability of the synthetic data used.

Drawing inspiration from the diverse successes observed with Adapters, encompassing both Adapter transfer and the innovative "Double Way Adapter Tuning" approach, we evaluated different alternative methodologies present in the literature. Our findings underscored the enduring efficacy of traditional Adapters as the most effective parameter-efficient method for enhancing children's ASR performance. We noted a tradeoff between accuracy and parameter efficiency. While some methods exhibited high parameter efficiency, they often resulted in significantly degraded results. Conversely, other approaches, while less parameter-efficient, yielded comparable or even superior results to entire model fine-tuning. To mitigate this tradeoff, we introduced a novel approach leveraging the inherent redundancy within the Feed-Forward Network components of Transformer-based models. Our “Shared-Adapters” methodology, wherein a single Adapter is shared across all layers instead of being allocated one per layer, demonstrated remarkable performance by outperforming entire model fine-tuning. Despite facing minimal score degradation compared to traditional Adapters, our shared adapter was trained on a substantially smaller number of parameters than any previous approach within this range of performance. They emerge as an outstanding candidate for achieving superior parameter efficiency transfer in the context of children's ASR, paving the way for more effective adaptation methodologies.