Towards improved children automatic speech recognition

Thesis

Thomas Rolland

Doctoral Program in Engenharia Informática e de Computadores

Supervisor: Prof. Alberto Abad

Examination Committee

Chairperson: Prof. Maria Luísa Torres Ribeiro Marques da Silva Coheur
Supervisor: Prof. Alberto Abad
Member of the Committee: Dr. Mathew Magimai Doss

FEBRUARY 2024

Acknowledgments

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam non luctus lectus. Nullam at ligula sed libero semper gravida. Duis vitae commodo ex. Quisque pharetra eget sapien vel rhoncus. Nullam viverra diam purus, a egestas est semper sit amet. Morbi consectetur, elit in vestibulum lobortis, dolor nisi ultricies orci, id porttitor nunc mi id ligula. Suspendisse vitae erat turpis. Vivamus vitae ligula sit amet est ultricies fermentum quis ac nisl. Donec consequat sagittis tellus malesuada varius. Quisque a euismod lacus. Integer dictum sem suscipit, fermentum odio tincidunt, tempus lectus. In hac habitasse platea dictumst. Nam aliquet iaculis lacus a sodales.

Sed imperdiet at ex et iaculis. Sed ac enim ex. Praesent purus nulla, tristique vel mi ornare, iaculis sollicitudin leo. Praesent ornare sem a lacus feugiat, eget ullamcorper ligula viverra. Phasellus venenatis in purus viverra maximus. Fusce id nisl semper, ultrices nisi ut, pharetra lacus. Nulla eget sapien mauris. Mauris vitae ligula nec lacus consequat hendrerit ut in nisi. Praesent vitae velit a purus consectetur interdum eget sit amet ex. Quisque lacinia egestas lorem tempor hendrerit. Vivamus molestie, diam et vestibulum consequat, purus quam luctus lectus, vitae porttitor odio sem non elit. Praesent tincidunt ante ac sodales rhoncus.

Morbi dignissim lobortis elementum. Nulla facilisi. Aenean ut risus eros. Nam feugiat faucibus massa, ut posuere odio. Maecenas finibus pharetra tellus, ut pharetra augue interdum at. Morbi vitae finibus lectus. Praesent pulvinar elit ut nunc varius, laoreet venenatis purus tristique. Nullam urna risus, condimentum quis suscipit porttitor, placerat vel neque. Nam suscipit, ante viverra laoreet ullamcorper, tellus nisl placerat urna, vitae semper risus metus in odio. Fusce ut efficitur lacus.

Abstract

In recent years, Automatic Speech Recognition (ASR) technology has advanced significantly, opening avenues for novel applications targeting young speakers. Potential use cases include innovative man-machine interactions through voice, automatic reading tutors, and speech pathologist assistants among others. However, the unique challenges presented in children's speech, characterised by high variability in both acoustics and linguistics components, impede the recognition performance of existing adult ASR systems.

This thesis focuses on improving ASR capabilities specifically tailored for children, aiming to develop a more robust and accurate system calledapable of handling the variabilities in speech across different age groups. To this end, various knowledge transfer approaches were explored.

The investigation initiated with knowledge transfer on Hybrid speech recognition models, employing transfer and multi-task learning strategies, and their combination. Subsequently, an end-to-end granular fine-tuning method was introduced to nhance and understand the adaptability of ASR systems to the nuances of children's speech during transfer learning. In parallel, Adapter transfer was examined alongside other parameter-efficient transfer techniques in order to discover a parameter efficient transfer method. Additionally, a novel approach involving data augmentation through synthetic data was explored to further enhance generalisation to children's speech patterns.

This research makes a significant contribution to the field of children's speech technology, providing a deeper understanding of knowledge transfer processes and introducing innovative approaches. The outcomes pave the way for improved human-computer interactions in educational, entertainment, and assistive technology applications specifically tailored for children. These results pave the way for future advances in ASR technology specially designed for the unique characteristics of young speakers.

Keywords

Automatic Speech Recognition, Children speech, Knowledge Transfer, Text-to-speech, Parameter efficiency

Resumo

Nos últimos anos, a tecnologia de Reconhecimento Automático da Fala (RAF) tem avançado, abrindo portas para novas aplicações destinadas a jovens falantes. Potenciais casos de uso incluem interações inovadoras homem-máquina através da voz, tutores de leitura automática e assistentes de patologia da fala. Contudo, os desafios únicos apresentados pela fala das crianças, com uma elevada variabilidade nas componentes acústica e linguística, dificultam o desempenho dos atuais sistemas RAF para adultos.

Esta tese foca-se na melhoria das capacidades de RAF adaptadas às crianças, com o objetivo de desenvolver um sistema mais robusto e preciso capaz de lidar com as variabilidades da fala em diferentes grupos etários. Foram exploradas várias abordagens de transferência de conhecimentos.

A investigação começou com a transferência de conhecimentos sobre modelos híbridos de reconhecimento da fala, utilizando estratégias de transferência e de aprendizagem multitarefa, assim como a sua combinação. Posteriormente, introduziu-se um método de afinação granular de ponta a ponta para melhorar e compreender a adaptabilidade dos sistemas RAF às nuances do discurso das crianças durante a aprendizagem por transferência. Em paralelo, a transferência Adapter foi examinada juntamente com outras técnicas de transferência eficientes em termos de parâmetros, a fim de descobrir um método eficiente em termos de parâmetros. Adicionalmente, foi explorada uma nova abordagem que envolve o aumento de dados através de dados sintéticos para melhorar ainda mais a generalização aos padrões de fala das crianças.

Esta investigação contribui significativamente para o domínio da tecnologia de fala para crianças, proporcionando uma compreensão mais profunda dos processos de transferência de conhecimento e introduzindo abordagens inovadoras. Os resultados abrem caminho para melhores interações homem-computador em aplicações educativas, de entretenimento e de tecnologia assistiva especificamente concebidas para crianças, apontando para futuros avanços na tecnologia RAF desenvolvida para as características únicas dos jovens falantes.

Palavras Chave

Reconhecimento automático da fala, fala infantil, transferência de conhecimentos, texto para voz, eficiência dos parâmetros

Contents

List of Figures

List of Tables

Acronyms

AD	Alzheimer's disease
AM	Acoustic Model
ASR	Automatic Speech Recognition
BERT	Bidirectional Encoder Representations from Transformers
CE	Cross-Entropy
CER	Character Error Rate
CMU	Carnegie-Mellon University
CNN	Convolutional Neural Network
CTC	Connectionist Temporal Classification
DCT	Discrete Cosine Transform
D	Deletions
DNN	Deep Neural Network
DRAFT	Domain Responsible Adaptation and Fine-Tuning framework
DWAT	Double-Way Adapter Tunning
ETLT	Extended Trentino Language Testing
fbanks	filterbanks
FFN	Feed-Forward Network
FFT	Fast Fourier Transform
G2P	Grapheme-to-Phoneme

GAN Generative Adversarial Network

GELU Gaussian Error Linear Unit

GLU Gated Linear Unit

GMM Gaussian Mixture Model

GMM-UBM Gaussian Mixture Model - Universal Background Model

GPT Generative Pre-trained Transformer

H/ASP Half / Attentive Statistics Pooling

Hz Hertz

HMM Hidden Markov Model

HMM-DNN Hidden Markov Model-Deep Neural Network

HMM-GMM Hidden Markov Model-Gaussian Mixture Model

I Insertions

KB Knowledge-Based

LF-MMI Lattice-Free Maximum Mutual Information

LM Language Model

LPC Log Power Spectrum

LSTM Long Short-Term Memory Network

MAP Maximum A-Posteriori

MFCC Mel-frequency cepstral coefficient

MHA Multi-Head Attention

MHSA Multi-Head Self-Attention

MLLR Maximum Likelihood Linear Regression

MLP Multi-Layer Perceptron

MLTL MultiLingual Transfer Learning

MLTL-olo MultiLingual Transfer Learning one-language-out

MTL	Multi-task learning
MyST	My Science Tutor
NLL	Negative Log-Likelihood
NLP	Natural Language Processing
OSA	Obstructive Sleep Apnea
PASE	Problem-Agnostic Speech Encoder
PD	Parkinson's disease
PER	Phone Error Rate
PETL	Parameter-Efficient Transfer Learning
PLP	Perceptual Linear Prediction
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
Seq2Seq	Sequence-to-Sequence
SDC	System Development Corporation
S	Substitutions
SAT	Speaker Adaptive Training
SCL	Speaker Consistency Loss
SGD	Stochastic Gradient Descent
SLP	Speech and Language Pathologist
SLT	Speech and Language Technologies
SSC	Spectral Subband Centroid
SSF	Scale and Shift Features
SSL	Self-Supervised Learning
SVM	Support Vector Machine
STT	Speech-to-text

SUR	Speech Understanding Research
SVD	Singular Value Decomposition
TDNN	Time-Delay Neural Network
TDNN-F	Factorised Time-Delay Neural Network
TL	Transfer Learning
TPA	Two Parallel Adapters
TSA	Two Serial Adapters
t-SNE	t-Distributed Stochastic Neighbor Embedding
TTS	Text-to-Speech
UAR	Unweighted Average Recall
UER	Unit error rate
VAD	Voice Activity Detection
VAE	Variational AutoEncoder
VITS	Variational Inference with adversarial learning for end-to-end Text-to-Speech
VTLN	Vocal Tract Length Normalisation
WER	Word Error Rate
WFST	weighted Finite-State Transducer

1

Introduction

Contents

1.1 Context

The faculty of expressing or describing thoughts, feelings and needs by using language is a fundamental capability in our daily lives. As per by the Oxford dictionary, *language is defined as the system of communication in speech and writing that is used by people of a particular country or area.* Consequently, language can be conceptualised as an intricate and rule-governed system that empowers individuals to convey abstract concepts, share experiences, and take part in nuanced forms of communication. Effective communication through language begins with the conceptualisation of the message to be transmitted (*conceptualisation*), followed by the selection of appropriate lexical items and subsequent grammatical encoding, culminating in their meaningful organisation (*formulation*). Subsequently, the linguistic representation is transformed into sound through the transmission of this representation from the brain to the muscles of the complex speech system including lips, larynx, glottis, lungs, jaw and tongue (*articulation*) [9]. Unlike language, which encompasses both spoken and written forms, speech specifically refers to the spoken manifestation of language.

The capability to speak and comprehend language is not inherently present but rather develops gradually over time with experience. Babies instinctively engage in pre-linguistic communication, using gestures, facial expressions, and vocalisations to articulate their basic needs. As language acquisition progresses, children transition to the babbling stage, experimenting with sound patterns. Eventually, speech emerges, marking a crucial milestone in communication development, and typically, children reach specific language milestones at particular ages. For example, around 12 to 18 months, a child usually utters their first words and starts imitating sounds. By the age of 4 to 5, children tend to formulate sentences and grasp more intricate concepts. Regarding speech sounds, younger children, approximately 1 year old, can produce basic speech sounds like */p/*, */b/*, */m/* while older children, around 5 years old, can articulate more complex sounds such as */r/* and */th/*. This developmental stage is referred to as language acquisition, and it plays a crucial role in a child's overall development. Indeed, our daily dependence on social and communication skills endures throughout our entire lives. Consequently, it is imperative for children to develop the capacity to interact effectively with others to achieve seamless integration into society across all aspects of their lives.

Regrettably, a subset of children experience speech disorders stemming from congenital conditions such as cleft palate, cerebral palsy, and prelingual deafness. Alternatively, certain individuals may acquire speech-related issues during childhood, encompassing cognitive developmental delays, breathing-feeding-swallowing disorders and traumatic brain injuries. Notably, in 2012, empirical data [10] highlighted that 7.7% of children aged 3 to 17 in the United States of America exhibited communication disorders, with 5.0% of this cohort specifically presenting speech-related problems.

Furthermore, findings [11] suggest that individuals with childhood speech disorders may confront an increased prevalence of mental health challenges, diminished social well-being, and reduced academic accomplishments in comparison to their healthy peers. This highlights the complex nature of speech disorders in children, the consequences of which extend into adolescence and adulthood. Hence, early identification and intervention play a pivotal role in mitigating the enduring effects on these children's social interactions, society integration, communication skills, educational progress, and overall well-being.

Pediatric Speech and Language Pathologists (SLPs) play a crucial role in providing therapy to help children overcome the effects of speech disorders and offer early diagnosis. The therapy typically includes exercises and assessments, which can be based on perceptual speech evaluations or standardised tests. To effectively engage children, these activities are often presented as games, taking into consideration the inherently limited attention span of children. Notably, SLPs frequently maintain long-term follow-ups with their patients, allowing them to monitor the evolution of speech quality over time and personalise exercises to the specific needs of each child. The adoption of this individualised therapeutic approach is essential for helping children achieve improved speech and communication skills.

However, a prominent challenge arises concerning the accessibility and availability of speech therapy services. Numerous children, particularly those residing in underserved or remote areas, encounter obstacles in accessing speech therapy resources. Additionally, the hospital environment introduces an additional layer of stress for children. While clinically necessary, therapeutic settings may inadvertently contribute to heightened anxiety and discomfort, as children may perceive it as intimidating. Furthermore, the logistical challenges associated with frequent hospital visits impose a substantial financial and time burden on families.

Another obstacle pertains to the continuity and consistency of therapy. Children may experience interruptions in their therapeutic journey due to factors such as financial constraints, scheduling conflicts, or alterations in healthcare coverage. These disruptions have the potential to impede progress and undermine the effectiveness of the therapy. Lastly, it is imperative to acknowledge that, despite professional training, inter-expert variability in perceptual assessments may persist, resulting in disparate diagnostic conclusions. To address these challenges, adopting a hybrid approach that combines in-person therapy with technology holds potential benefits [12, 13]. Teletherapy, for instance, has emerged as a promising avenue to bridge geographical gaps and deliver therapy services remotely [14].

In this context, Speech and Language Technologies (SLT) have emerged as highly pertinent within the domain of speech therapy [15]. These technologies encompass a spectrum of computational tools designed to analyse, understand and provide objective and precise automated assessments. Another benefit lies in their potential integration into gamification frameworks, thereby augmenting children's involvement during therapy [16]. Moreover, the ability to record speech utter by the patient during a session using SLT enables post-session thorough analysis and long-term monitoring by the therapist. Due to the aforementioned reasons, the development of such tools has gained considerable attention, empowering patients to engage in exercises beyond therapy sessions, notably in a home setting. Several

Figure 1.1: Illustrated herein are some examples of children's Speech and Language Technology applications that were developed during the course of this thesis. On the left is a running platformer game, where the user's voice controls the character. Pitch dictates running and jumping actions, while energy modulates the velocity of these actions. On the right, a reading task game is depicted, wherein a robot instructs the user to read designated words.

SLT examples developed within the scope of this thesis are illustrated in Figure 1.1.

1.2 Problem statement

Recent years have seen increased integration of SLT into various aspects of our daily lives, impacting a wide range of environments, including homes, transportation, education, and even the military. Noteworthy examples encompass voice assistants, hands-free computing, healthcare systems, automatic helplines, and speech-to-speech translation services. The performance progress in these applications was made possible through the use of machine learning techniques, especially deep learning approaches, the increasing computational capacities of our devices, and the ever-growing volume of data available to train and improve these systems.

Children represent a promising target audience for SLT due to the inherent complexities of conventional computer interfaces, which pose challenges for them and limit their capacity to fully benefit from digital platforms. Indeed, children commonly face difficulties in manipulating mouse and keyboard inputs. Additionally, the abstract nature of traditional man-machine interfaces can impede the understanding necessary for effective interaction. In this context, speech-based systems emerge as a promising alternative, offering a more natural and accessible means for children to interact with technology. Through the use of speech recognition technologies, these speech systems mitigate the barriers associated with conventional interfaces, providing a fluid and intuitive interaction paradigm that aligns more closely with the developmental stages and cognitive abilities of young users. Another potential application of SLT for children could be in the development of automatic reading tutors. Given that the process of learning to read is individual and varies for each child, a personalised automatic reading tutor could assist in tailoring the learning experience to the specific needs of each student. This has the potential to reduce the workload on teachers and provide additional support in the crucial skill of reading acquisition. Finally, the growing presence of voice assistants in home settings underscores the importance of reliable Automatic Speech Recognition (ASR) for children. In this context, a reliable ASR system becomes crucial to ensure a positive and effective user experience of seamless interaction with voice-enabled devices in various home environments.

As previously mentioned, SLT applications are gradually making their way into the field of atypical speech, particularly for children. While these automatic tools are currently in their early stages and have limitations, there is indeed a rising interest in implementing atypical speech and language therapy systems with a focus on assisting SLPs. In this context, systems capable of automatically recognising speech content, assessing pronunciation quality and automatically detecting speech pathologies could be highly valuable in supporting pediatric SLPs and patients.

All of these objectives require the implementation of a robust ASR system specifically tailored for healthy children, serving as a foundational model. Nevertheless, while speech recognition technologies for adult speakers have made substantial advancements, leading to increased accuracy, the performance of ASR systems for children remains underperforming in comparison. This discrepancy results in unreliable systems for children and, by extension, their use in SLT applications. This diminished performances can be attributed to a combination of factors, including intra- and inter-speaker variability, limited linguistic and phonetic knowledge, and the scarcity of available data.

In this thesis, we will undertake a comprehensive investigation into the intricacies of children's speech, closely examining the inherent differences between children and adults in the domain of ASR. Through this examination, the objective is to analyse the constraints associated with the application of adult-based systems to children's speech and, subsequently, to outline the methodologies present in the literature for enhancing ASR systems. The overarching goal of this thesis is to establish a robust foundational system that effectively addresses the recognition of children's speech.

Initially, the context of this thesis aimed at addressing pathological speech for children. However, due to constraints in data availability, specifically the absence of a meaningful pathological children's speech dataset, experiments related to pathological speech in children were not included in this thesis. The shift in focus towards improving ASR for healthy children was motivated by the importance of establishing a solid ASR foundational system focusing on children. This shift broadens the potential applications of our research, ranging from personal assistants, automatic reading tutors and voice interactions with computer interfaces. Despite the pivot towards healthy children's speech, this thesis encompasses an investigation into adult pathological speech detection, as detailed in Annex A.

Our work specifically aims to answer the following research questions:

1. Which knowledge transfer approach is best for efficiently modelling and improving automatic recognition of children's speech? Can these approaches be used to efficiently exploit low-resource children's speech data from multiple languages?

2. How do end-to-end automatic speech recognition models achieve state-of-the-art results for children's ASR when finetuned from an adult model? Particularly, what are the components that are most important to fine-tune?

3. Is it possible to develop a parameter-efficient automatic speech recognition model for children? Can we further improve the parameter efficiency with other architectures?

4. Is it possible to use children's synthetic speech to extend the amount of children's data? How can we control the quality and speakers' variability?

1.3 Contributions

This thesis commenced with a comprehensive exploration of the current state-of-the-art in children's Automatic Speech Recognition (ASR). We present an examination of the fundamental determinants that contribute to the decline in ASR performance for children's speech. Additionally, we meticulously assess current research on children's speech. The primary aim was to identify potential avenues for improvement throughout the course of this thesis.

Subsequent to the literature review, we start our research by the implementation of Hidden Markov Model-Deep Neural Network (HMM-DNN) models for children ASR. We explored different strategies to reduce the gap observed between children and adults in the context of both English and European Portuguese speech. We identified the effectiveness of knowledge transfer methods, specifically, transfer learning and multi-task learning. Transfer learning adapts speech recognition adult models by fine-tuning their weights for children's speech. On the other hand, multi-task learning exposed models to both adult and children's speech datasets simultaneously during training. In an innovative synthesis, we proposed to combine transfer learning and multi-task learning into a unified approach, the multi-task transfer learning framework. We applied this approach to multiple low-resource children's datasets from diverse language sources resulted in a publication at LREC 2022:

- **Rolland, Thomas**, Alberto Abad, Catia Cucchiarini, and Helmer Strik. "Multilingual Transfer Learning for Children Automatic Speech Recognition." *Language Resources and Evaluation Conference* (2022).

Thereafter, our research turned into the end-to-end paradigm, motivated by the encouraging improvements observed in the end-to-end children's ASR performance. We introduced a novel detailed transfer learning approach, called Partial fine-tuning, where our objective was to gain a thorough understanding of the specific components within the end-to-end architecture that significantly contributed to the remarkable improvements in recognition scores. The identification of the most relevant components allowed the development of specific algorithms aimed at further improving the model's recognition performances.

Particularly, we explored the integration of an additional set of parameters, the Adapters, directly into the original ASR model. This integration facilitated a parameter-efficient approach to adapt the model, accepted at ICASSP 2024:

- **Rolland Thomas** and Alberto Abad. "Exploring adapters with conformers for children's automatic speech recognition." *International Conference on Acoustics, Speech and Signal Processing* (2024).

In response to the scarcity of large children's speech datasets, we delved into the exploration of leveraging synthetic speech to augment the existing dataset. However, our investigation revealed that a mismatch between real and synthetic data hindered the results. To address this challenge, we introduced additional processing steps to efficiently incorporate synthetic data. We proposed a double-way approach, wherein the synthetic data underwent an additional set of parameters. This innovative methodology contributed to an enhanced ASR system tailored for children and was published at ICASSP 2024:

- **Rolland Thomas** and Alberto Abad. "Improved children's automatic speech recognition combining adapters and synthetic data augmentation." *International Conference on Acoustics, Speech and Signal Processing* (2024).

In tandem with the primary focus of enhancing children's ASR, this thesis extends its scope to the detection of pathologies directly from speech. This secondary investigation retains relevance within the broader context of the thesis, particularly as we initially aimed to address the specific needs of children with pathological speech. We explored the use of embedding extracted from pre-trained models for the detection of different pathologies such as Alzheimer's disease, Parkinson's disease, obstructive sleep apnea and COVID-19:

- Anna Pompili, **Thomas Rolland**, and Alberto Abad. "The INESC-ID multi-modal system for the ADReSS 2020 challenge." *Interspeech* (2020).
- Catarina Botelho, Francisco Teixeira, **Thomas Rolland**, Alberto Abad, and Isabel Trancoso. "Pathological speech detection using x-vector embeddings." *arXiv preprint* arXiv:2003.00864 (2020).
- Rubén Solera-Ureña, Catarina Botelho, Francisco Teixeira, **Thomas Rolland**, Alberto Abad, and Isabel Trancoso. "Transfer Learning-Based Cough Representations for Automatic Detection of COVID-19." *Interspeech* (2021).

1.4 Structure for the thesis

The structure of this thesis comprises eight chapters. In Chapter 2, we first establish the context of the thesis and understand the challenges associated with automatic children's speech recognition. This chapter also provides an overview of the automatic speech recognition systems, along with an examination of the latest approaches specifically tailored to address the unique challenges posed by children's ASR. Furthermore, a compilation of children's speech corpora is presented.

Following, in Chapter 3, we present our work on the hybrid speech recognition paradigm, focusing on the evaluation of different knowledge transfer approaches and their combinations. In Chapter 4, we shift towards the end-to-end paradigm, evaluating the role of the different components of the ASR model during transfer learning and proposing our partial fine-tuning procedure. Subsequently, in Chapter 5, we validate the use of Adapters as a parameter-efficient knowledge transfer approach for children's ASR. In Chapter 6, we build upon this knowledge to propose a novel approach to use imperfect synthetic data as data augmentation during transfer learning. Next, in Chapter 7, we investigate possible alternative approaches for parameter-efficient transfer learning, introducing a novel method using a shared Adapter across the different layers of the model. Finally, in Chapter 8, we conclude the thesis with a summary of the different works and results studied in this thesis, as well as some perspectives of future work.

2

Background - Children automatic speech recognition

Contents

ASR, or Speech-to-text (STT) refers to the process of mapping a raw spoken audio utterance into its corresponding text. The potential use of STT applications across diverse fields has motivated the need for robust and reliable ASR systems. These applications extend across various sectors, encompassing academia, medicine, industry, and the military. Notably, ASR has made significant progress in recent years, thanks to the attention and investment from both industry and public authorities. This support has resulted in the deployment of applications such as voice assistants, hands-free interfaces, medical assistance, live translation, and more, all of which are widely used and accepted today. Nowadays, the majority of ASR applications are mainly developed and optimised for adult speech. Demonstrating high performance in conditions close to those encountered during the training phase. This focus on adult speech is explained by the immediate potential of ASR applications for this target audience. In addition, training STT models on adult speech has both advantages of data availability and relatively stable aspects of adult speech characteristics making the training easier. Indeed, adult speech is often more standardised, with established linguistic conventions and stable features. However, the challenge arises when such systems are applied to recognise speech in mismatched scenarios, like atypical speech such as accented, pathological or children's speech. For example, in the context of children's speech, ASR algorithms often exhibit a decline in performances, frequently two to five times worse [17]. In this context of children, this difference in performance can be mostly attributed to the intra- and inter-speaker variability. In fact, speech serves as a channel not only for linguistic content but also for paralinguistic cues that reveal aspects of the speaker's identity, including age, gender, state of health, emotional state and regional origin. Although this additional layer of information is highly valuable for human-to-human communication, it introduces a new level of complexity, making the development of a reliable ASR system more challenging [18]. Furthermore, several external factors further negatively impact the performance, including noise, speaker variability, mispronunciation, and the quality of the recording [19, 20].

The potential applications of STT in education and entertainment have led to a growing interest in ASR for children. Indeed, children represent a demographic group that could well benefit from such applications for a number of reasons. Firstly, the complexity of traditional computer interfaces, such as keyboards and mice, can pose problems for young children, making speech interfaces a more accessible and user-friendly option. Secondly, speech and language applications, including reading tutors and speech and language acquisition assistants, promise to address educational inequalities among children and facilitate their integration into society by giving them personalised and tailored attention.

In this chapter, we first present the various challenges associated with children's speech recognition. These challenges encompass the unique characteristics of children's speech, including high acoustic and linguistic variability, as well as the limited amount of labelled data available for training. Then, we provide a brief introduction to ASR, tracing its historical development from early pattern recognition approaches to the advent of statistical models and the contemporary move towards end-to-end models. This historical background provides an understanding of the underlying principles behind ASR technologies. Subsequently, we review the state-of-the-art methods specifically designed to address the challenges posed by children's ASR. The aim of this in-depth exploration is to provide a clear overview of the different techniques that are being used to improve children's speech recognition. Finally, the chapter concludes with a discussion that synthesises the different perspectives presented earlier and the ones selected for this thesis.

2.1 Children speech recognition challenges

In this section, we explore the distinct challenges posed by children's speech to ASR systems. In particular, we will explore the main differences with adult speech. Indeed, the divergences between child and adult speech are mainly due to the continuous growth and intellectual development of children. This growth has direct repercussions on automatic speech recognition scores. In order to present the different challenges associated with ASR for children, we have identified at least three of the main factors that degrade recognition performance. First, we examine the acoustic variability of children's speech. The acoustic characteristics of children's speech differ considerably from those of adults due to factors such as vocal tract size, pitch modulation and articulatory differences. These variations represent a significant challenge for speech recognition systems, which are often trained on adult speech datasets and have never encountered such variations. Taking this acoustic variability into account becomes imperative for the development of accurate and robust ASR models adapted to the unique characteristics of children's speech. Next, we will present the linguistic and phonetic knowledge inherent in children. Indeed, children's language evolves dynamically over time, with vocabulary expansion, phonetic development and language mastery. This linguistic evolution also raises challenges for ASR systems, as they need to be robust to age-specific linguistic variation and imprecise pronunciation. In a similar way as the acoustic variability, effective modelling of these linguistic variabilities is important for children's speech recognition. Finally, we present the challenge posed by the limited availability of corpora of children's speech. Unlike adult speech, data corpora containing labelled examples of children's speech are relatively rare and small in size. This scarcity constrains the training of robust ASR models, as it limits their exposure to the various linguistic and acoustic variations of children's speech.

2.1.1 Speech variability

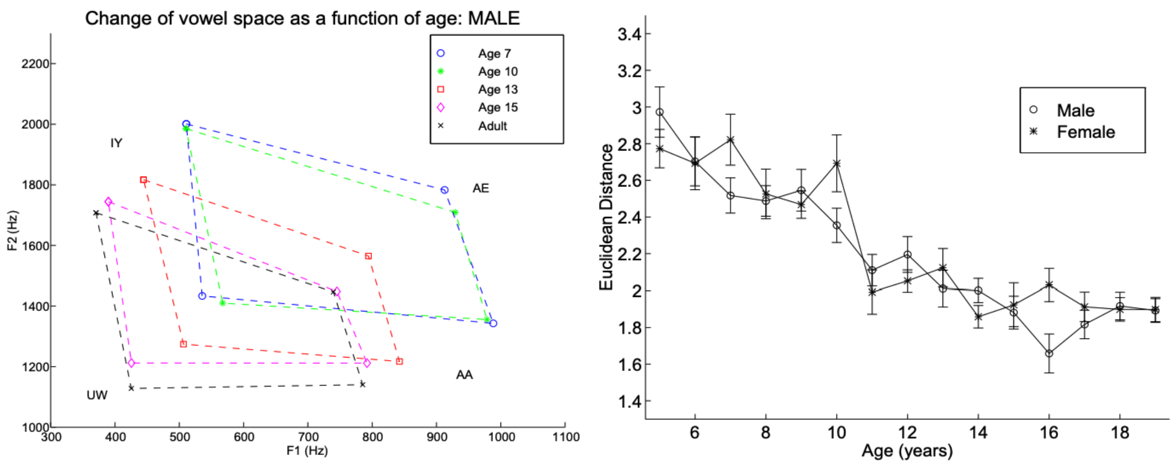
Speech production is a complex process involving the synchronised actions and collaboration of several elements of the speech production apparatus. These include the vocal cords, tongue, lips and mouth. The coordination of these elements leads to fluctuations in air pressure, producing a wave called speech. Speech is therefore essentially a measure of air pressure over time. The waveforms of human speech encompass a range of frequency components from 20 Hertz (Hz) to 20 kHz. These are detected and processed by the auditory system and the human brain. Because speech is based on frequency components, an accurate understanding of these frequency components, such as fundamental frequency and formant frequencies, is essential for the development of reliable speech processing tools.

The fundamental frequency, often called F0, plays a crucial role in the analysis of speech signals. It characterises the (quasi-) periodic average oscillations produced by the vibrations of the vocal folds. Measured in Hz, F0 is often considered to be the acoustic correlate of pitch. F0 shows an inverse relationship with the vibrating mass of the vocal cords, leading to distinct F0 values for different demographic groups. As a general rule, adult males have lower F0 values, ranging from around 100 to 150 Hz. Women, on the other hand, tend to have higher F0 values, generally between 200 and 300 Hz. Children, whose vocal cords are smaller, often have even higher F0 values, generally ranging from 300 to 450 Hz. These variations in F0 contribute to the perceptual differences in pitch between individuals of different ages and genders. According to [2], significant differences in F0 between male and female speakers generally appear from the age of 12. For male speakers, decreases in F0 were observed on average between the ages of 11 and 15, at the time of puberty, and did not change significantly after the age of 15. Furthermore, it was observed that the relatively large variation between male subjects at the ages of 13 and 14 also implies that the starting point of puberty varies among speakers in these age groups. For female speakers, the pitch drops between the ages of 7 and 12, and there is no significant change in pitch after this age. In addition, the change in F0 in female subjects is more gradual as compared to male speakers.

It is essential to emphasise that F0 is not a static parameter; on the contrary, it exhibits continuous variation within a sentence. This dynamic nature allows F0 to be used expressively in speech, conveying nuances such as accent, emotion and intonation patterns. Variations in F0 help to distinguish between different types of speech acts, such as statements, questions and exclamations.

A formant frequency refers to a concentration of acoustic energy centred around a specific frequency in a speech waveform. As defined by the Acoustical Society of America, it is *"a range of frequencies in which there is an absolute or relative maximum in the sound spectrum. The frequency at the maximum is the formant frequency"*. Formants play a crucial role in characterising vowel sounds and distinguishing between them. In speech analysis, the first three formants, known as F1, F2 and F3, are commonly used for their importance in capturing the acoustic characteristics of vowels and their contribution to the timbre of speech sounds.

The pioneering study by Peterson and Barney in 1952 [21] marked a turning point in the exploration of the formant components of vowels, particularly in the context of children's speech. Researchers undertook a comparative analysis, examining the vowel frequencies of children and comparing them with those of adult men and women. This research was the first to show significant variations in vowel frequencies based on the speaker's age and gender. Building upon this foundational work, subsequent studies [1, 2, 22] have provided further insights into the acoustic characteristics of children's speech. These investigations have consistently demonstrated a correlation between acoustics and children's age, attributing these variations primarily to the growth of the children's vocal apparatus. The scaling behaviour of formant frequencies with respect to age is shown in Figure 2.1(a). Here, the evolving vowel space, defined by four reference vowels (*/IY/*,*/AE/*,*/AA/* and */UW/*) linearly decreases with age and aligns with the adult level around the age of 15. Additionally, as highlighted in [1], the vowel space becomes more compact as age increases, indicative of a downward trend in the dynamic range of formant values. These variations and age-related differences emphasise the critical challenge of inter-speaker variability, especially for young children.



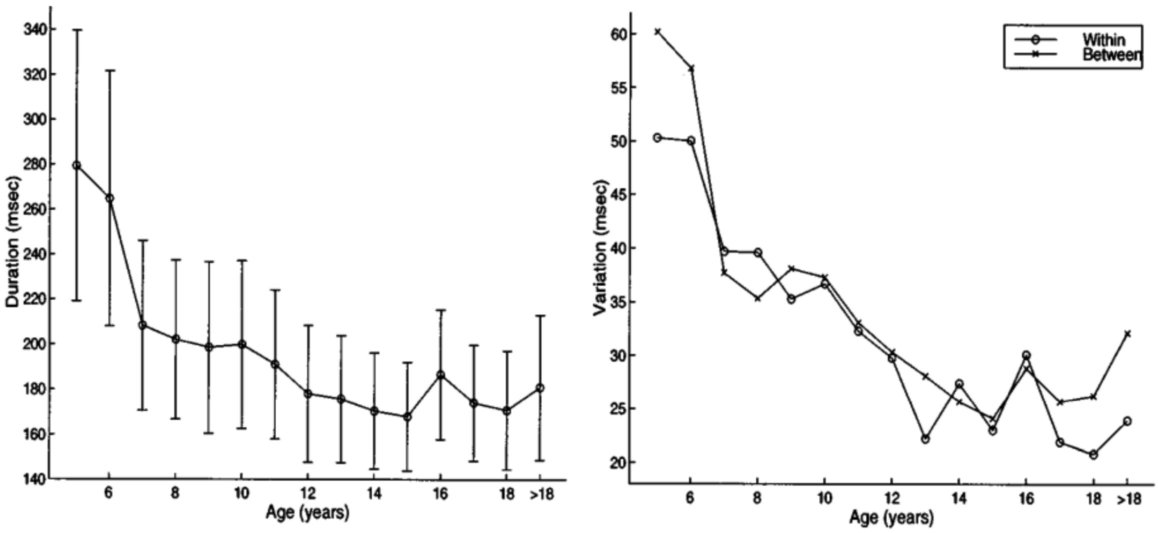
(a) Changes in F1-F2 vowel space as a function of age

(b) Mean cepstral distance between the two repetitions of the same vowels

Figure 2.1: Formant and cepstral variability. Figures taken from [1]

In addition to inter-speaker variability, [2] also highlight the fact that children's speech also exhibits intra-speaker variability, signifying that the speech produced by the same individual can exhibit variations. This variability arises from two primary sources. Firstly, as previously discussed, the acoustic characteristics of children can significantly differ at different ages due to the ongoing growth of their vocal apparatus. Secondly, even at the same age, the same child may produce variable speech, even when articulating the same vowel. As depicted in Figure 2.1(b), the average cepstral distance between two repetitions of the same vowels by the same child tends to decrease with age, particularly after the age of 10. This reduction in intra-speaker variability is attributed to the progressive mastery of speech articulation components as children grow and mature their motor skills abilities. The decrease in cepstral distance suggests a more coherent and standardised articulation of vowels over time.

Segmental duration is another important aspect of human speech. A segment, as defined by Crystal [23], is: *"Any discrete unit that can be identified, either physically or auditorily, in the stream of speech"*. These segmental durations could be of vowel or sentence duration. Vowel duration, in particular, is



(a) Averaged-vowel duration across all vowels and subjects in each age group

(b) Within- and between-subject variations. The between-subject variation is reduced by a factor of 2.0

Figure 2.2: Segmental duration variability. Figures taken from [2]

of significant importance in vowel discrimination. Research, as presented in [2], investigates how these characteristics change in children's speech. As demonstrated in Figure 2.2(a), the average vowel duration in children exhibits variations with age. On average, younger children tend to have longer vowel durations, resulting in a slower speaking rate. However, as children become more comfortable with the processes of speech production with age, vowel duration gradually decreases. Similarly to children frequency variations, segmental duration exhibits intra-speaker variability, as illustrated in Figure 2.2(b).

In conclusion, dealing with both intra- and inter-speaker variability in children's speech, particularly in those under the age of 15, poses a substantial challenge for developing high-performance speech processing models. Especially, this challenge is exacerbated in the context of children's ASR where the age of the child is often unknown. In addition, the dynamic processes of the vocal tract growth, changes in linguistic knowledge and the maturation of control of speech apparatus occur simultaneously and overlap, making it considerably more challenging to accurately disentangle and model their effects. The intricate nature of children's speech, marked by intra and inter-speaker variations, underscores the necessity for sophisticated and adaptive models that can accommodate the unique characteristics of these speakers.

2.1.2 Language and phonetic knowledge

Language is a complex and multifaceted system of communication that involves the use of symbols, such as words to convey meaning. It is a unique human ability and serves as a fundamental aspect of human cognition and social interactions. Linguists have identified five basic components of language [24], including phonology (sounds), morphology (structure and construction of words), semantics (meaning), syntax (grammar and sentence structure), and pragmatics (how language is used in context). It allows individuals to express thoughts, share information, and engage in social interactions. It is important to note that, languages vary across cultures and regions, exhibiting a rich diversity of sounds, structures, and expressions. Additionally, language can be spoken, written, or signed, and evolves over time. For children, the mastery of language is a crucial milestone in their cognitive development. Furthermore, language plays a central role in shaping culture, identity, and the transmission of knowledge to them. The children's ability to use language develops with age, achieving adult capabilities around the age of 13, as indicated by [2]. This progression enables the children to transition from producing simple sounds and words to generating more complex sounds and fully articulated sentences.

During the process of language acquisition, children, constrained by their limited linguistic knowledge, often make pronunciation errors and encounter disfluencies [25]. According to [26], these errors may include a variety of phenomena, such as:

- **Substitution**: Involves the inadvertent replacement of the correct pronunciation of an entire word with an alternative word.

- **Omission**: Refers to the act of leaving out or neglecting a part of speech, a word, or a phrase that would typically be included in a grammatically correct or complete sentence.

- **Mispronunciation**: Involves the act of pronouncing a part of a word incorrectly, deviating from the standard or expected pronunciation of this word.

- **Pause and Hesitation**: Entails temporary breaks or delays in speech during which a speaker might refrain from producing sound or articulate speech in a hesitant manner.

- **Filler and mumbling**: Filler encompasses linguistic elements used during pauses or hesitations when a speaker needs time to think, including unintelligible sounds, words, or phrases without significant meaning. Mumbling is characterised by unclear or indistinct speech, often marked by low volume, unclear articulation, and imprecise pronunciation.

- **False-start**: Refers to an instance where a speaker begins a sentence or a word and then stops abruptly before completing it. This interruption is often followed by a restart or a correction to articulate the intended message more accurately.

- **Sound-out**: Involves a pronunciation strategy in which a speaker articulates a word by pronouncing each sound or phoneme separately, rather than blending them together seamlessly.

Potamianos and Narayanan's study [27] revealed significant insights into the variability and characteristics of children's linguistics. They found that inter-speaker variability is approximately twice as much as intra-speaker variability. Additionally, their research found that the rate of mispronunciations is twice as high for children aged 8 to 10 compared to those aged 11 to 14. Conversely, the trend is reversed for filler and pauses, where the older group exhibits a higher rate. Furthermore, younger children, of 8 to 10 years, tend to produce more false-starts and breathing.

In adult speech, pronunciation errors and disfluencies are also present, but their occurrences are typically lower than what is observed in the speech of children, as supported by studies such as [28, 29]. In addition, these studies used language models specifically trained on children's speech, demonstrating their advantages over the use of adult language models. These findings underscored the differences between children's and adults's linguistics, encompassing variations in grammatical structures as well as the presence of mispronunciations and disfluencies. Such insights are crucial for the development of effective language models tailored to the unique characteristics of children.

2.1.3 Data scarcity

In recent years, the emergence of deep learning has brought significant advancements in the ASR field. The combination of increased computational power and the availability of large datasets has played a pivotal role in these improvements. The success of deep learning is largely attributed to Deep Neural Network (DNN)s, which can approximate complex non-linear functions. With the help of this capability, DNN excels in capturing complex patterns and accurate representations of speech data. However, the efficacy of a DNN in capturing speech patterns depends a lot on the availability of training data. Indeed, using large-size datasets is pivotal for enhancing the capabilities and generalisation of DNN-based ASR systems. Notably, top-performing ASR systems like Whisper have been trained on exceptionally large datasets, surpassing 680,000 hours of data collected from the web [30]. There is a noticeable trend in the speech research community towards the collection of larger datasets, exemplified by initiatives such as the LibriSpeech dataset, which comprises around 1,000 hours of speech [31], and the GigaSpeech dataset, featuring 10,000 hours of speech [32].

Unfortunately, despite rare recent efforts to collect larger children datasets [33–35], the majority of publicly available children corpora include fewer than fifty hours of speech. This is significantly less than a typical adult speech corpora, which usually contains hundreds or even thousands of hours of data. Furthermore, the majority of the accessible children's data are English corpus [33, 35–38]. However, English is a resource-rich pluricentric language which should be seen more as an exceptional case, rather than an average representative. A compilation of existing datasets containing children's speech will be presented in 2.4.

The scarcity of children's speech datasets availability can be partially attributed to a combination of ethical, legal, and technical challenges. Collecting speech data from children raises ethical concerns related to obtaining consent, ensuring privacy, and protecting minors. The heightened awareness of online safety and security concerns further complicates the creation and sharing of datasets that include children's speech, as there is a need to safeguard against potential misuse and ensure the anonymity of participants. Beyond ethical and legal considerations, technical challenges also play a role. Children's speech patterns, language development, and pronunciation can vary significantly across different age groups, as explained in previous sections, making it more challenging to create datasets that accurately represent the diversity of children's speech. Moreover, the resource-intensiveness of collecting high-quality speech data from children, which involves careful planning, recruitment efforts, and coordination with schools or parents, can further contribute to the limited availability of such datasets. Finally, collecting speech data from children is a challenging and time-consuming task. Various factors can significantly impact the quality of the gathered speech. These include children's short attention spans, recording environments that might be noisy (such as classrooms), and the quality of the speech, which is highly dependent on the task at hand (reading tasks are generally more complex for children).

The importance of having a large database of children's speech to cover different variabilities has been emphasised in a study conducted by Liao in 2015 [39]. In this work, the researchers trained an ASR model using a large in-house corpus of children's speech. Notably, this corpus was comparable in size to typical adult speech corpora. The result was the attainment of state-of-the-art performance by the ASR model, even demonstrating competitiveness with adult speech recognition systems. This study underscores the crucial role of large-size and diverse children's speech datasets in developing robust and high-performance ASR models tailored to the unique characteristics of children's speech.

2.2 Introduction to automatic speech recognition

In this section, a brief historical overview of ASR is presented, laying the foundation for a subsequent exploration of predominant trends and modules within ASR systems. This comprehension is necessary for the following sections of this thesis. While not exhaustive, this overview provides essential insights, with certain topics falling beyond the scope of this thesis are intentionally omitted. For a more exhaustive understanding of ASR, readers are encouraged to consult references such as [40–42]. This section is structured as follows: Firstly, we present the historical evolution of ASR, followed by a description of traditional HMM-based ASR systems, succeeded by an explanation of the end-to-end paradigm. Concluding this section, a discussion on automatic speech recognition metrics is presented.

2.2.1 A brief history of Automatic Speech Recognition

2.2.1.A Early Days

The origins of speech recognition technology can be traced back to the 1950s and 1960s, with initial projects focusing on isolated word recognition with a speaker-dependent system. One of the earliest projects in this direction was the creation of a digit recogniser at Bell Telephone Laboratories in 1952. This recogniser demonstrated the automatic recognition of telephone-quality digits spoken at normal speech rates by a single adult male speaker, achieving an impressive accuracy of up to 99%. The system relied on formant frequency approximations to recognise entire words. It is important to underscore that within this recogniser, there was an absence of explicit modeling of syllables, consonants, vowels, or any other sub-word units. In this recogniser, a word was treated as a single unit, which was then compared with ten standard digit patterns to find the best match. The recognition process first involved extracting two frequency ranges, below and above 900 Hz. Motivated by the observation that these two frequency ranges approximately align with the frequencies of the first two formants in speech. Then, these formant approximations were plotted on a 2D plot with a trace interruption period of 10ms. Finally, when presented with new audio, the system generated a new plot, compared it to the reference plots of the ten digits, and returned the closest match by computing the highest relative correlation coefficient. Figure 2.3 illustrates an example of the 2D representation of the ten digits.

In 1962, IBM developed Shoebox, a device capable of recognising 16 spoken words, including the ten digits and command words such as "plus," "minus," and "total." This system employed a similar pattern-matching algorithm to the one used in the Bell Telephone Laboratories's recogniser.

Nevertheless, extending such a system to a larger vocabulary would be impractical. Indeed, the template-matching approach necessitated saving each word representation on disk and comparing the unknown spoken word with all these representations. Therefore, when attempting to scale this approach to automatic large vocabulary recognition, issues of time and disk usage complexity emerged as a significant

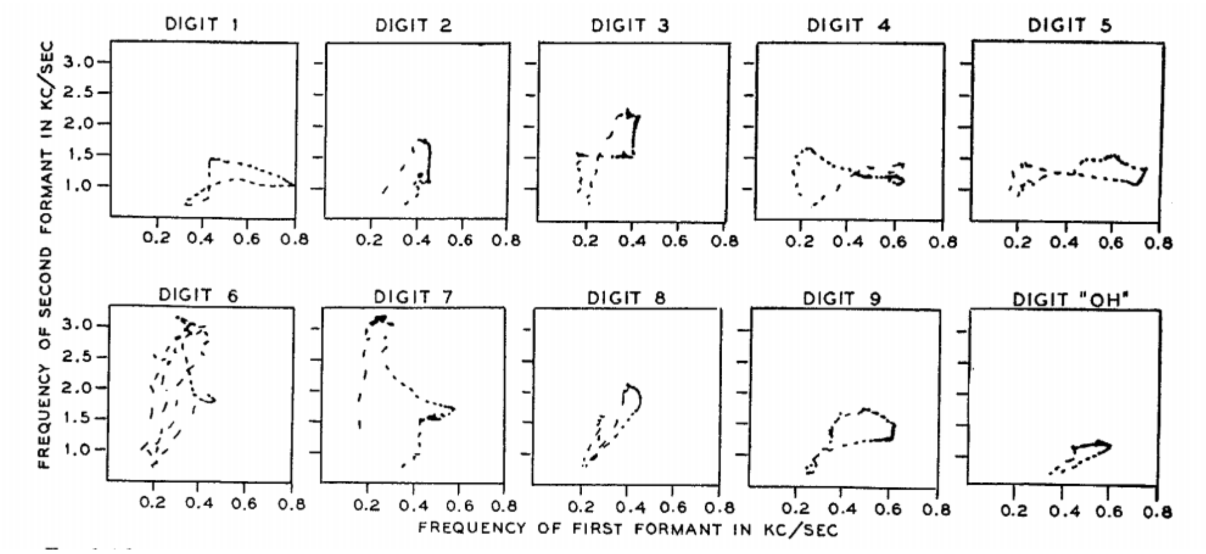


Figure 2.3: Example of a standard digit pattern from Davis et al. 1952

challenge. Furthermore, in order for the circuit to deliver an accuracy of the same range for a new speaker, a preliminary analysis of the speech of that individual and subsequent circuit adjustments were necessary. These limitations underscored the need for more scalable and adaptive approaches and led the field of automatic speech recognition to continue to evolve.

2.2.1.B The Speech Understanding Research program

In the early 1970s, subsequent to the initial success of pattern-matching algorithms in single-word recognition, the Advanced Research Program Agency of the U.S. Department of Defense, ARPA, initiated funding for a five-year program called Speech Understanding Research (SUR). The overarching goal of SUR was to "obtain a breakthrough in speech understanding capability that could then be used toward the development of practical man-machine communication system". Within the context of this program, four distinct research groups were funded: two from Carnegie-Mellon University (CMU), one from Bolt Beranek and Newman Inc. (BBN Hwim), and the last one from System Development Corporation (SDC). Each group was assigned a specific task, such as dealing with facts about ships, travel budget management, and document retrieval. The ultimate objective for each group was to create a system capable of recognising simple sentences within the context of their assigned task, using a vocabulary of 1,000 words and achieving a Word Error Rate (WER) of 10% in a reasonable amount of time.

The realisation that the pattern-matching word identification strategy could not be directly applied to the challenge of sentence understanding prompted a redesign of the single-word identification system. On the first hand, one key recognition was that the acoustic characteristics of words can vary considerably based on the context of the sentence. The impracticality of storing each word and all its possible different variations on disk became apparent. Moreover, determining the boundaries of each word was almost an impossible task, and even if these boundaries were identified, the pattern-matching computation, involving comparisons with each of the 1,000 stored words and all their possible variations, would be time-consuming and exceed the reasonable time requirement. Secondly, another crucial consideration in the redesign of the system was that the length of the spoken sentence is variable and unknown, in contrast to the relatively fixed length in single-word identification tasks.

To address these challenges, a shift was made to a smaller unit than the word for modelling speech -namely, phonemes. Phonemes are the smallest distinctive and meaningful units that compose speech. Each language is associated with a finite set of phonemes, typically fewer than 50, which can be combined to form words. This shift enabled a more efficient and flexible representation of speech, accommodating the variability in the pronunciation of words.

Among all the systems proposed in the project, the Harpy system implemented by Lowerre in 1976 by the CMU team exhibited the best performances [3]. Harpy is a speaker-specific system that uses a pattern-matching algorithm at the phoneme level instead of the word level. The system employs a set of 98 phonemes and diphones a pair of consecutive phonemes-, encompassing pronunciations of all words, along with a graph compiling all accepted sentences using 15,000 states. When a new spoken utterance is provided to the system, it undergoes an initial processing phase, involving low-pass filtering at 5 kHz, digitisation at 10,000 samples per second, and computation of 14 linear prediction coefficients with a 10ms shift. To speed up the decoding process, analogous adjacent acoustic segments are grouped together. Subsequently, these audio segments are compared against the 98 phoneme templates, and the system deduces the optimal path over the decoding graph. Figure 2.4 provides an example of a decoding graph in the Harpy system.

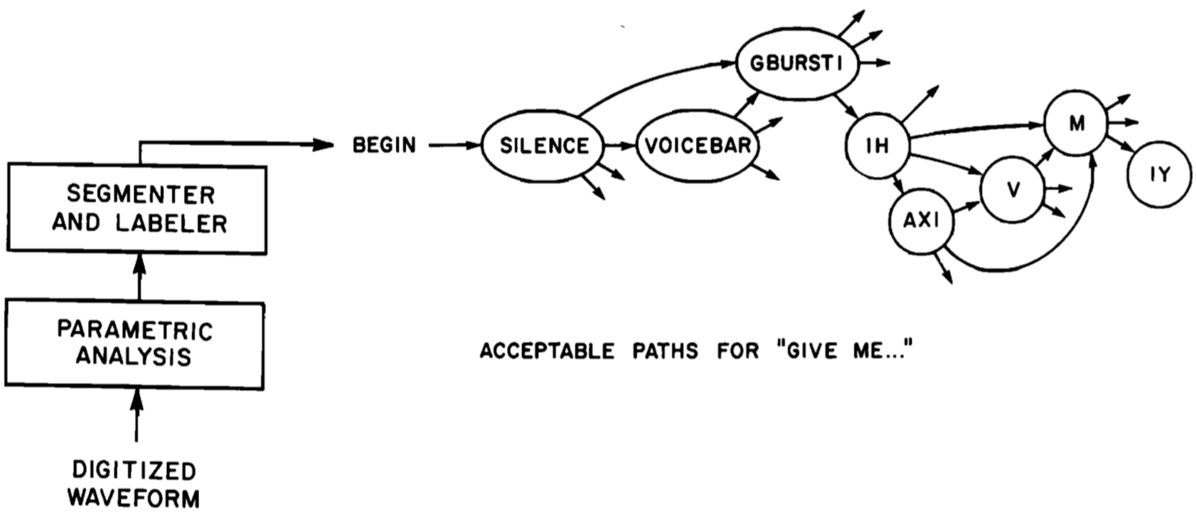


Figure 2.4: Example of a decoding graph in the Harpy system for the sentence “GIVE ME” from [3]

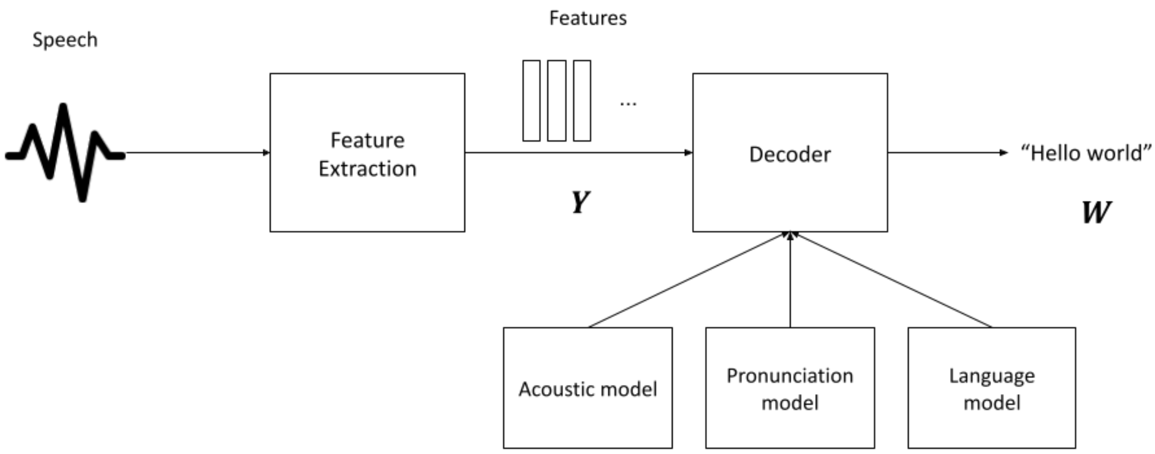


Figure 2.5: Architecture of a HMM-based speech recognition system

Notwithstanding the achievements and success of the Harpy system, it has limitations that hinder its broader applicability. As a speaker-specific system, it requires tuning for each new speaker over the 98 phoneme templates. Additionally, the system is constrained to recognise a vocabulary of no more than 1,000 words and relies on simple handcrafted grammar, making it less reliable for handling spontaneous speech. Moreover, the decoding time of the system falls short of real-time requirements. These constraints highlight the need for more generalisable and efficient speech recognition systems, especially for handling diverse speakers and spontaneous speech scenarios. Therefore, as research progressed, the limitations of pattern-matching-based approaches became apparent. This realisation prompted the exploration of probabilistic modelling techniques, marking a shift towards more sophisticated and adaptable approaches in ASR.

2.2.2 Traditional automatic speech recognition systems

In the 1970s, the introduction of Hidden Markov Models (HMMs) led to a paradigm shift in ASR research, moving away from traditional pattern-matching methods towards statistical modelling [43]. Indeed, HMMs are particularly effective at capturing the sequential and temporal nature of speech. They assume that speech can be represented as a sequence of hidden states, each state corresponding to a distinct phonetic unit. HMM models the transitions between these states and, at each state, generates observable acoustic features. The hidden aspect refers to the fact that the underlying states are not directly observed but inferred from the observable features. HMMs are particularly well suited to modelling speech dynamics, as they can represent the variability of speech sounds over time. In the context of ASR, HMMs have been widely used to model phonemes, words or sub-word units.

Building on this foundation, the 1980s saw the emergence of Gaussian Mixture Models (GMMs), which further enhanced the statistical modelling capabilities of ASR [44]. GMMs allowed for a more flexible representation of the probability distributions underlying speech features. GMMs are used to model the statistical distribution of acoustic features associated with each hidden state in an HMM. Commonly, a system that uses both HMM and GMM is referred to as a Hidden Markov Model-Gaussian Mixture Model (HMM-GMM) framework. By using GMM, they assume that the distribution of features can be approximated by a mixture of several Gaussian distributions. GMM are versatile in capturing the variability of speech sounds, allowing a more flexible representation of the acoustic units. In ASR, GMMs are commonly used to model the emission probabilities associated with each state in an HMM. This means that given a particular state, the GMM provides the likelihood of observing a specific set of acoustic features. By combining the temporal modelling capabilities of HMMs with the statistical representation power of GMMs, this framework effectively captures the complex relationship between acoustic features and phonetic units.

Finally, in the 1990s, statistical grammar also played a crucial role, providing a structured framework for incorporating linguistic information into ASR systems [45]. Statistical grammars represent a category of grammars that integrate statistical information to characterise the probability of diverse linguistic structures. In contrast to traditional rule-based grammars, which articulate a language's syntax through explicit rules, statistical grammars adopt a data-driven methodology. They assign probabilities to various linguistic constructions based on observed frequencies within a designated corpus.

The components illustrated in Figure 2.5 represent the traditional ASR pipeline. To this day, these components continue to form the core of modern HMM-based ASR systems. However, the recent evolution in the ASR field has been marked by a significant shift from the GMM to the adoption of DNN. Driven by DNN ability to effectively model complex patterns and hierarchies in speech data, DNNs have demonstrated superior performance, contributing to enhanced accuracy and efficiency in speech recognition systems [46]. Called hybrid models, or HMM-DNN, by effectively combining both the strengths of HMMs and DNNs. A DNN is a subtype of artificial neural networks consisting of multiple layers of interconnected neurons. These neurons, organised in layers, receive an input signal, and each connection between neurons is characterised by a weight that signifies its strength. In addition, each neuron is associated with a bias weight, providing an additional learnable parameter. During training, the network adjusts these weights and biases to minimise the difference between predicted and actual outputs, a process known as backpropagation. Moreover, a non-linear activation function is placed within neurons, as it enables the network to model intricate, non-linear patterns of the data. The weights, biases and non-linearity allow DNNs to capture complex relationships and representations from the data, learning hierarchical features and abstracting information across multiple layers of the network.

In this statistical framework, the continuous speech audio waveform is transformed into a sequence of fixed-size acoustic vectors, denoted as $\boldsymbol{X} = x_1, ..., x_T$. The goal of the ASR system is to determine the sequence of words, $\boldsymbol{w} = w_1, ..., w_L$, that is most likely to have produced the observed acoustic vector sequence $\boldsymbol{X}$. This is formulated as finding the word sequence $\hat{w}$ that maximises the conditional probability $P(\boldsymbol{w}|X)$. More formally:

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmax}}\{P(\boldsymbol{w}|X)\} \tag{2.1}$$

However, directly modelling the conditional probability $P(\boldsymbol{w}|X)$ can be challenging. Bayes' Rule offers a way to express this probability in terms of more manageable components, specifically by decomposing it into the product of the likelihood of the observed acoustic vector sequence given the word sequence $P(X|\boldsymbol{w})$ and the prior probability of the word sequence $P(\boldsymbol{w})$. Therefore, equation 2.1 became:

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmax}}\{\frac{P(X|\boldsymbol{w})P(\boldsymbol{w})}{P(X)}\} = \underset{\boldsymbol{w}}{\operatorname{argmax}}\{P(X|\boldsymbol{w})P(\boldsymbol{w})\} \tag{2.2}$$

Here, the likelihood $P(\boldsymbol{X}|\boldsymbol{w})$ is determined by the acoustic model component, capturing the probability of observing the acoustic vector sequence $\boldsymbol{X}$ given the word sequence $\boldsymbol{w}$. In parallel, the prior probability $P(\boldsymbol{w})$ is determined by the language model component. The term $P(X)$ is not essential for determining the maximum probability and can be omitted in the context of finding the most likely word sequence. Subsequent sections will provide a more in-depth exploration of these distinct components and their processes.

2.2.2.A Feature extraction

The first step in the STT pipeline, as depicted in Figure 2.5, is the feature extraction from the speech signal. The feature extraction component plays a crucial role in capturing pertinent information about the linguistic content of speech. In addition, the efficacy of speech recognition systems is intricately tied to the quality of the extracted features. To this end, for each time step, the continuous waveform is transformed into a small fixed-size vector. An acceptable assumption is that speech is considered stationary within the time span covered by a single vector. Consequently, feature vectors are typically computed at intervals of 10 milliseconds, often with a 25-millisecond overlapping window.

Within the domain of ASR, a broad range of different acoustic features can be employed. However, in the context of HMM-based models, the predominant features encompass Perceptual Linear Prediction (PLP), Melspec, filterbanks (fbanks), andMel-frequency cepstral coefficient (MFCC). Particularly, MFCCs as introduced by Davis and Mermelstein [47], stand as the predominant features in HMM-GMM and HMM-DNN architectures. The process of extracting MFCCs typically involves several steps to capture essential information from the speech signal. First, a preemphasis filter is applied to the signal. Subsequently, the signal is segmented into frames, and a Hamming window with a duration of 25 milliseconds is applied to each frame. The frames are then transformed into the frequency domain using the discrete Fast Fourier Transform (FFT), resulting in a magnitude spectrum. The next stage involves pass-

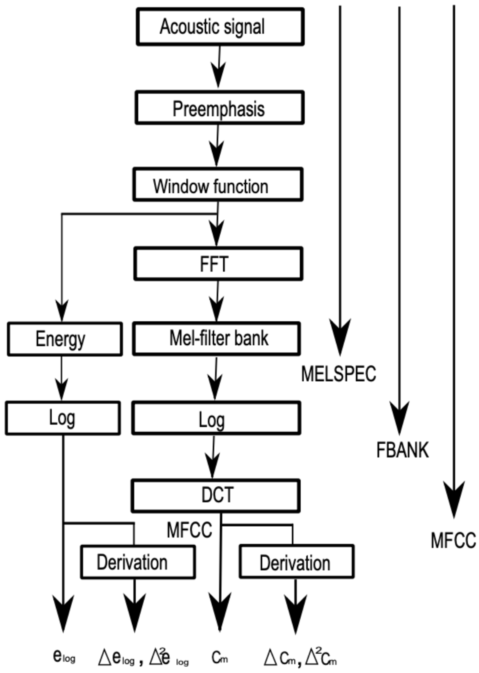


Figure 2.6: Principal block scheme of extraction of main speech features for ASR: Melspec, fbanks and MFCC coefficients from [4]

ing the magnitude spectrum through a bank of triangular-shaped filters. Extracting features at this point yields melspec features. The energy output from each filter is log-compressed, concluding the extraction process at this stage would result in fbanks features. Finally, MFCCs are obtained by transforming the filterbank features into the cepstral domain using the Discrete Cosine Transform (DCT) to decorrelate the energies obtained from the filterbanks. The overall representation of this extraction process is illustrated in Figure 2.6.

To incorporate information about the dynamics of the speech signal, the feature vector for each time step is augmented with the first and second-order derivatives, commonly denoted as Δ (Delta) and $\Delta\Delta$ (Delta-Delta), respectively. The first-order derivative coefficients, often referred to as Δ coefficients, are calculated by taking the difference between consecutive feature vectors. Mathematically, the Δ coefficients for a feature vector at time t are computed as follows:

$$\Delta_i = \frac{\sum_{n=1}^{N} n(f_{i+n} - f_{i-n})}{2\sum_{n=1}^{N} n^2} \tag{2.3}$$

Here, f_i represents the feature at the instant i.Typically, n is set to 2, indicating that the first-order derivatives are calculated by considering the differences between the feature at the current time t and its neighbouring features at $t \pm 2$. The $\Delta\Delta$ coefficients, also written Δ^2, represent the second-order derivatives and are computed in a similar manner as Δ in equation 2.3 by taking the difference between

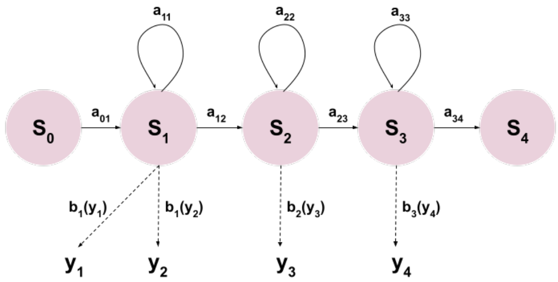


Figure 2.7: Three-state Hidden Markov Model for modelling phones

consecutive Δ coefficients instead of the spectral feature f. The concatenation of the first-order derivative (Δ) and second-order derivative (Δ^2) features with the spectral features is denoted as x_i. Mathematically, this concatenation can be expressed as follows:

$$\boldsymbol{x_i} = [f_i \qquad \Delta_i \qquad \Delta_i^2] \tag{2.4}$$

Here, f_i represents the spectral feature at the instant i, Δ_i represents the first-order derivative feature at the same instant, and Δ_i^2 represents the second-order derivative feature at the same instant. The resulting feature vector x_i encapsulates information about the spectral content of the speech signal as well as its temporal dynamics, providing a more comprehensive representation for subsequent processing by the ASR system, especially the acoustic model.

2.2.2.B Acoustic model

The role of the Acoustic Model (AM) is to determine $P(\boldsymbol{X}|\boldsymbol{w})$. While employing a classifier such as GMM models with one GMM per phone is a straightforward approach, it tends to disregard the temporal dependencies inherent in speech, such as co-articulation. Indeed, accurately categorising each frame necessitates the consideration of not only the current frame but also its context, encompassing both previous and following frames. Additionally, there are acoustic differences at the beginning, middle, and end of each phone, which further complicate the classification task. To address these concerns, the HMM framework has been proposed as a solution [48]. HMMs offer temporal flexibility, incorporating concepts such as self-looping, and provide a well-understood framework with effective learning (Expectation Maximisation) and decoding (Viterbi) algorithms.

In the HMM terminology, the observed variables, denoted as y_i, correspond to the acoustic features (e.g., speech signal), while the hidden variables are represented as states, denoted as s_i. The states are generated using a first-order Markov process, where the i^{th} state s_i depends solely on the previous state s_{i-1}. The transition from one state to another is determined by the transition probability a_{ij}.

Upon entering a state s_i, an observation in the form of an acoustic vector is emitted, and this emission is modelled by the distribution $b_i(\cdot)$ associated with that state. Typically, this emission distribution is modelled by a GMM. It is assumed that all observations are independent given the states that generated them. A fundamental HMM configuration for speech recognition, represented in Figure 2.7, involves a three-state model representing the beginning, middle, and end of a phoneme, along with an initial and final state. This model, known as a monophone HMM-GMM, constitutes a basic unit for phonetic modelling. For example, since English has 44 phonemes [49], a monophone system in English will have 44 separate HMM-GMM. However, due to the influence of co-articulation effects and the desire to capture phonetic variations based on context, more complex models, such as triphone systems, are employed [50]. A triphone system aims to model each phoneme in its specific phonetic context, leading to a significantly larger number of required models. Indeed, for a language with N phonemes, there should be N^3 models to train. For example, in English, which has 44 phonemes, the total number of models would be $44*44*44$ resulting in 85,184 models. To manage the computational complexity, these models are often clustered using decision trees [51]. This hierarchical clustering helps capture phonetic variations efficiently while working with limited available data.

The concept of hybrid models gained prominence in the 1990s with the integration of Multi-Layer Perceptron (MLP) as replacements for GMMs in the HMM-GMM system [52, 53]. Subsequently, the introduction of DNNs, which are basically MLPs with a large number of hidden layers, in 2012 marked a significant advancement in various ASR tasks [46]. The efficacy of DNNs lies in their capability to capture complex and highly non-linear relationships between inputs (e.g., audio features) and outputs (e.g., phoneme labels) due to the substantial number of parameters induced by the deep architecture.

However, the training of HMM-DNN and HMM-GMM models differs. Neural networks necessitate labelled data for training, which includes both input features and corresponding output labels (e.g., phoneme labels). Standard speech training data often lacks this detailed labelling, providing only audio waveforms and utterance transcriptions. Consequently, the training of HMM-DNN models relies on alignments generated by an HMM-GMM. Typically, the training process of HMM-DNN involves initial flat-start monophone training with HMM-GMM, followed by iterative steps into triphone training with more precise alignments and subsequently the DNN training. In consequence, the precision of the HMM-GMM alignment directly impacts the efficacy of DNN model training. As ASR continues to advance, the integration of various DNN architectures, including Convolutional Neural Networks (CNNs) [54], Long Short-Term Memory Networks (LSTMs) networks [55], and Time-Delay Neural Networks (TDNNs) [56], further refines the modelling of spatial and temporal relationships, laying the foundation for more sophisticated and context-aware speech recognition systems.

Phoneme	Example	Translation
AA	odd	AA D
AE	at	AE T
AH	hut	HH AH T
AO	ought	AO T
AW	cow	K AW
AY	hide	HH AY D
B	be	B IY
CH	cheese	CH IY Z
D	dee	D IY
DH	thee	DH IY
EH	Ed	EH D
ER	hurt	HH ER T
EY	ate	EY T
F	fee	F IY
G	green	G R IY N
HH	he	HH IY
IH	it	IH T
IY	eat	IY T
JH	gee	JH IY
K	key	K IY

Phoneme	Example	Translation
L	lee	L IY
M	me	M IY
N	knee	N IY
NG	ping	P IH NG
OW	oat	OW T
OY	toy	T OY
P	pee	P IY
R	read	R IY D
S	sea	S IY
SH	she	SH IY
T	tea	T IY
TH	theta	TH EY T AH
UH	hood	HH UH D
UW	two	T UW
V	vee	V IY
W	we	W IY
Y	yield	Y IY L D
Z	zee	Z IY
ZH	seizure	S IY ZH ER

Figure 2.8: Phoneme set and examples of CMU dictionary using 39 phonemes from [5]

2.2.2.C Pronunciation model

The pronunciation model in ASR, often referred to as a dictionary or lexicon, plays a crucial role in establishing the correspondence between phonetic units, such as phonemes, and the respective words in the language. Indeed, words are essentially comprised of phonetic segments, and the pronunciation model specifies how these segments combine to articulate the pronunciation of each word. It is noteworthy that a single word may have multiple pronunciations. This mapping takes the form of an entry where all possible words are associated with their corresponding sequence of phones. Examples of words along with their corresponding phonetic sequences are illustrated in Table 2.8. Traditionally, this mapping is obtained manually, relying on phonetic and linguistic knowledge.

Furthermore, the integration of statistical Grapheme-to-Phoneme (G2P) tools [57] augments the lexicon by facilitating the generation of pronunciations for words that may not be explicitly included in the dictionary.

2.2.2.D Language model

The Language Model (LM), often referred to as grammar, holds a pivotal role in ASR, responsible for determining the probability $P(\boldsymbol{w})$ of equation 2.2. Beyond its use in ASR, the applications of language models extend into diverse fields including Natural Language Processing (NLP) [58], computational biology [59], and data compression [60]. The two most successful approaches to language modelling widely adopted in ASR are, respectively, statistical methods and models based on deep learning.

Statistical LMs rely on traditional techniques like HMM and N-grams. Particularly, N-grams, are the simplest approach for language modelling, they estimate the likelihood of the next word based on the context of the preceding n words as follows:

$$P(\boldsymbol{w}) = P(w_1, w_2, \ldots, w_L) = \prod_{i=1}^{L} P(w_i | w_{i-n}, \ldots, w_{i-1}) \tag{2.5}$$

The level of context can vary from the case of $n = 1$ -the 1-gram -or unigram- which considers each word independently to higher-order n-grams that incorporate more extensive context for enhanced accuracy. The unigram model would be defined as follows:

$$P(\boldsymbol{w}) = P(w_1, w_2, \ldots, w_L) = \prod_{i=1}^{L} P(w_i) \tag{2.6}$$

Despite the evident advantages of employing a larger n for enhanced contextual information in language modelling, practical considerations and computational limitations often impose constraints on the choice of n in real-world ASR applications. The escalating combinational complexity associated with higher n values becomes computationally demanding, presenting challenges for efficient processing, storage, and training. As a result, the majority of ASR applications typically use trigrams or 4-grams, striking a balance between contextual accuracy and computational feasibility.

Furthermore, determining the start of sequence probability precisely introduces intricacies, especially with larger n-grams. Additionally, the reliance on training data poses a notable limitation for n-grams, particularly in estimating the likelihood of unseen words. This deficiency becomes apparent when facing vocabulary expansion or encountering out-of-vocabulary terms, necessitating specific techniques such as smoothing to address these challenges. [61].

In contrast, deep learning-based LMs has opened up a new era, employing neural networks with complex architectures to achieve remarkable modelling capabilities. Unlike traditional n-grams, these models demonstrate a high degree of flexibility, and ease training and do not require as many resources as n-grams to be efficient. Recent advances in language modelling, exemplified by state-of-the-art models such as Bidirectional Encoder Representations from Transformers (BERT) [62] or Generative Pre-trained Transformer (GPT) [63], are built on deep learning networks.

A key factor contributing to the success of deep-learning language models is the incorporation of attention mechanisms. Unlike the limited contextual awareness of n-grams, attention mechanisms allocate varying degrees of importance to different words within a sentence. This approach enables the model to focus more on crucial elements, capturing intricate dependencies and semantic information that contribute to a more accurate language representation. The attention mechanism's ability to discern and prioritise important words enhances the overall performance and effectiveness of deep-learning language models.

2.2.2.E Decoder

In the context of ASR, the decoder role is to use the language, acoustic, and pronunciation models to determine the most likely word sequence, denoted as $\hat{\boldsymbol{w}}$, given a corresponding sequence of acoustic features, denoted as $\boldsymbol{Y}$ (as referred in equation 2.1). This is achieved by employing dynamic programming to search through all potential sequences. Notably, the Viterbi algorithm [64] is instrumental in efficiently solving this decoding problem. However, in practical applications, a direct implementation of the Viterbi algorithm could become challenging, especially for continuous speech, where considerations such as model topology, language model constraints, and computational constraints must be taken into account. N-gram language models and cross-word triphone contexts further complicate the search space. To address these challenges, various approaches have emerged.

One approach involves constraining the search space by maintaining multiple hypotheses in parallel [65] or dynamically expanding it as the search progresses [66]. Another alternative is to use beam search where the idea is to prune search paths which are unlikely to succeed. More recently, recent advancements in weighted Finite-State Transducer (WFST) technology offer a comprehensive solution by integrating all necessary information, including acoustic models, pronunciation, and language model probabilities, into a single, highly optimised network [67, 68]. This approach provides both flexibility and efficiency, making it a valuable tool for ASR. As demonstrated by the Kaldi speech recognition toolkit [69], it stands out as a widely adopted toolkit that leverages WFSTs for decoding.

Although decoders are primarily designed to find the best solution to the aforementioned probability computation in equation 2.1, they can also generate a set of the N-best hypotheses. This capability enables multiple passes over the data without incurring the computational expense of repeatedly solving the probability computation from scratch. The word lattice [70] serves as a convenient structure for storing these hypotheses, consisting of nodes representing points in time and spanning arcs representing word hypotheses. Word lattices offer remarkable flexibility, allowing for rescoring by using them as input recognition networks. Furthermore, they can be expanded to facilitate rescoring by a higher-order language model.

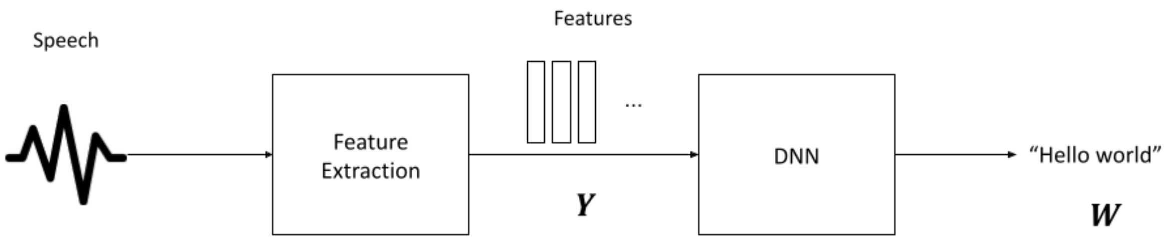


Figure 2.9: Architecture of an end-to-end speech recognition system

2.2.3 End-to-end automatic speech recognition

End-to-end speech recognition represents a paradigm shift in the field, presenting a streamlined and holistic approach compared to traditional HMMs-based systems. In contrast to conventional modular systems that incorporate distinct acoustic, pronunciation, and language models, end-to-end architectures aim to simplify the ASR process by directly mapping input audio signals to transcriptions within a single neural network model as illustrated in 2.9. Indeed, one of the key disadvantages of hybrid models is the factorised training of all modules independently, which can lead to error accumulation and mismatches between the different components. Therefore end-to-end strategy simplifies the overall system design, eliminating the requirement for pre-aligned training data and post-processing of outputs, thereby fostering a more data-driven and automatic learning process. In this paradigm, word-level transcriptions are transformed into character-level transcriptions. Considering the sequence of fixed-size acoustic vectors $\boldsymbol{X} = x_1, ..., x_T$ and the corresponding character sequence $\boldsymbol{Y} = y_1, ..., y_N$, where T and N represent the numbers of frames and the length of the character sequence respectively, the objective of end-to-end models is to learn the conditional probability of the character y_i given the input $\boldsymbol{X}$ and the preceding output $y_{<i}$:

$$P(Y|X) = \prod_{i=1}^{N} P(y_i|X, y_{<i}) \tag{2.7}$$

In recent years, with the growing interest in the end-to-end ASR paradigm, these systems have demonstrated increasing performances. Particularly, in scenarios where a large amount of labelled training data is available. Recently, the use of end-to-end systems has allowed comparable or even superior performance compared to traditional HMM-based systems across various ASR datasets [71,72]. However, despite its promising results, end-to-end speech recognition faces several challenges. Most importantly, these systems require a large corpus of training data to operate effectively. In the absence of such a corpus, the model may struggle to perform correctly [73]. In addition, handling rare or out-of-vocabulary words remains a significant challenge. Finally, the generalisation across new acoustic conditions of the model can be an issue. Therefore, there is a need for research on end-to-end speech recognition, particularly to simplify the end-to-end design, training, and robustness of these speech recognition systems.

The overall structure of end-to-end systems depicted in figure 2.9 highlights the integration of the acoustic model, pronunciation model, and language model into a single neural network in end-to-end architectures. Nevertheless, the feature extraction stage remains identical to traditional ASR systems. In end-to-end systems, the most commonly employed fixed-size acoustic features are filterbanks. Motivated by the higher flexibility that they provide and their capability of capturing relevant information from the speech signal compared to MFCCs.

Transitioning to the end-to-end paradigm has necessitated the development of new training approaches. Indeed, training an end-to-end model differs significantly from training traditional HMM-based systems. Notably, two training procedures have emerged in the literature of end-to-end ASR: Connectionist Temporal Classification and sequence-to-sequence architectures. Each approach comes with its own distinct features and advantages, and the subsequent sections offer a more in-depth exploration of these two methodologies. It is worth mentioning that while these are distinct methods that can function independently, they can also be employed in conjunction.

2.2.3.A Connectionist Temporal Classification

The first step towards end-to-end ASR was made with the introduction of the Connectionist Temporal Classification (CTC) objective function, by Grave et al. [74]. The main innovation of CTC is that it eliminates the need for pre-segmented training data, enabling the model to automatically learn the alignments between the N input speech frame $\boldsymbol{X}$ and the output sequence of T phones $\boldsymbol{Y}$ if $N \leq T$, representing a departure from the traditional HMM-based models.

To this end, the CTC objective function consists of two essential sub-processes: path probability calculation and path aggregation. Consider $\mathcal{V}$ as the set of possible paths of phone-label sequences of length T, and let p_k^t denote the probability of observing the label k at time t. It is noteworthy that CTC necessitates the length of the label sequence Y to be equal to T. To address any length difference between N and T, a blank label "-" is introduced, representing the probability of observing no label.

First, the path probability calculation involves computing the conditional probability of any path $\pi \in \mathcal{V}$ given the observed acoustic features $\boldsymbol{X}$.Mathematically, this is expressed as:

$$p(\pi|X) = \prod_{t=1}^{T} p_{\pi_t}^t, \forall \pi \in \mathcal{V} \tag{2.8}$$

Where π_t denotes the label at time t in sequence path π. Considering all possible paths and their respective probabilities is crucial, but direct computation becomes infeasible due to the exponential number of potential paths.

To address the computational challenges, the path aggregation step comes into play. Its purpose is to sum the probabilities of paths that correspond to the same label sequence $\boldsymbol{Y}$ by marginalising over all possible paths. The path aggregation also merge the same contiguous labels and deletes the blank label. For example, two different paths "b-ii-r-d" and "b-i-r-dd" become "bird". This is mathematically represented as:

$$p(Y|X) = \sum_{\pi \in \theta_Y} p(\pi|X) \tag{2.9}$$

Where θ_Y is a subset of $\mathcal{V}$ of all possible path π corresponding, after aggregation, to the label sequence Y.

2.2.3.B Sequence to sequence

The Sequence-to-Sequence (Seq2Seq) architecture, initially proposed by Sutskever for machine translation [75] stands as an important paradigm of end-to-end ASR systems. The original context of its application was for machine translation tasks where word sequences were translated from one language to another. The inherent challenge lies in the differing lengths of input and output sequences. However, this architectural framework, especially with the integration of attention mechanisms [76], has showcased remarkable versatility, extending its efficacy across diverse applications such as image captioning [77], conversational modelling [78], text summarisation [79], and ASR [80].

The core components of a standard Seq2Seq model consist of an Encoder and a Decoder modules. The Encoder processes input sequences of variable length, transforming them into a sequence of vectors often denoted as the "internal state" or "hidden representation". This sequence of vectors encapsulates the crucial information extracted from the input features. Subsequently, the Decoder uses this sequence of vector representation to generate an output sequence of tokens iteratively. Mathematically expressed as:

$$p(y_1, ..., y_T) = \prod_{i=1}^{T} p(y_i|y_0, ...y_{i-1}, f(H)) \tag{2.10}$$

where $f(H)$ represents a function of the Encoder's output $H = (h_1, ..., h_N)$. Notably, in Seq2Seq models incorporating attention mechanisms, $f(H)$ includes attention to selectively focus on relevant segments within H for predicting the current target token. The Seq2Seq objective function is formulated to train the model by maximising the conditional probability of generating the target sequence given the input sequence using Negative Log-Likelihood (NLL) loss or Cross-Entropy (CE) loss.

A significant difference from CTC-based models lies in the fact that Seq2Seq models do not make independent assumptions about output labels. Instead, they directly model the conditional probability of each target token given the preceding tokens in the output sequence and the encoder's output. This end-to-end approach empowers Seq2Seq models to handle sequences of varying lengths, making them particularly advantageous for speech recognition, where precise alignment between input and output is challenging.

2.2.4 Automatic Speech Recognition metrics

In the domain of ASR, metrics serve as indispensable tools for assessing the accuracy and effectiveness of systems. These metrics provide quantitative evaluations that act as a crucial benchmark, enabling researchers, developers, and engineers to objectively measure the performance of their ASR models. The evaluation process of ASR systems involves a meticulous comparison between system-generated transcriptions and reference transcriptions. Among these metrics, WER is the most commonly used for assessing STT systems. The ASR system's output word sequence is matched with a reference transcription, and the number of Substitutions (S), Deletions (D), and Insertions (I) are summed. As a result, WER is calculated as follows:

$$WER = \frac{S + D + I}{N} \times 100 \tag{2.11}$$

Where N is the total number of words in the reference transcription. As a result, a lower WER score is indicative of better system performance. The computation of WER is based on the Levenshtein distance, operating at the word level rather than the phoneme level. The primary goal is to quantify the dissimilarity between the ASR system's output and the reference transcription. Notably, a WER score greater than 100% can be attained when the number of mistakes surpasses N, while a score of 0% is the minimum achievable when there are no errors in the ASR hypothesis compared to the reference.

State-of-the-art ASR systems developed by leading research institutions and companies have achieved WER scores ranging from around 4.3% to 8.13% on well-resourced benchmark datasets such as the Switchboard corpus for conversational speech recognition [81] and the French subset of the read speech Common voice dataset [82] respectively. However, it is important to note that WER scores are task-specific, and some tasks are still achieving high WER scores, especially in challenging conditions or for certain languages and accents, such as 38.9% for the CHiMIE-6, a low resource noise speech dataset [83].

Beyond WER, there exist other metrics derived from the same fundamental equation but operating at different levels of transcription. Examples include Phone Error Rate (PER) based on phonemes and Character Error Rate (CER) which operates on character instead of word, a metric used to quantify Mandarin STT systems. These metrics provide a nuanced evaluation by concentrating on specific linguistic units, contributing to a comprehensive assessment of ASR system performance in diverse contexts.

2.3 Children automatic speech recognition

Addressing the challenges highlighted in Section 2.1 has prompted diverse initiatives across various segments of the ASR pipeline. This involves exploring improvements at the feature level, with the development of novel extraction techniques and adaptations. Data augmentation strategies have been employed to enrich training datasets, offering the model exposure to a more diverse range of children's speech patterns. Modifications in annotation detail have been explored, refining the labelling process to better capture the nuances of children's speech.

Beyond feature-level interventions, advancements in acoustic model structures have been pursued. This involves exploring new architectures and refining existing ones to better accommodate the characteristics of children's speech. In addition, innovative training procedures have been introduced to optimise model learning from the available data.

This section reviews the state-of-the-art for each of these aspects in more detail. Following this comprehensive review, we will identify and delineate the specific approaches that emerge as promising or particularly impactful for addressing the challenges associated with children's ASR. These identified approaches will serve as the focal points for the subsequent works of this thesis.

2.3.1 Feature extraction stage

The feature extraction stage is critical for identifying relevant speech signal components for both traditional and end-to-end ASR. This phase is characterised by the intentional elimination of speaker-dependent attributes, such as fundamental frequency, while simultaneously preserving the integrity of phoneme-dependent characteristics, notably exemplified by formant frequencies as described in section 2.2.2.A. In the context of children's speech recognition, the acoustic characteristics pose unique challenges, including close fundamental frequency and formant values, as well as phonetic class overlap due to formant variability. Moreover, research conducted by Ghai et al. [84, 85] showcased that the use of MFCCs, a feature commonly used in ASR, exhibits an acoustic mismatch compared to adult MFCCs. This mismatch is exacerbated due to inadequate smoothing of pitch-dependent distortion present in the speech of child speakers. To tackle these challenges, various strategies have been proposed to enhance acoustic features for children's speech, encompassing new feature extraction, feature adaptation, and the concatenation of additional features.

An early step in the direction of better feature extraction for children was the introduction of PLP features in 1990 by Hermansky [86]. PLP features demonstrated a more accurate representation of formants of children's speech. An additional strategy employed was the use of binary-weighting of MFCCs. This approach involves truncating some of the higher coefficients to remove those with non-sufficient smoothing, as proposed by [84]. The objective was to refine the representation of MFCCs by selectively retaining non-distorted coefficients. Furthermore, Gamma-tone filterbanks were employed to wrap the spectrum on a different scale, aiming to decrease variance compared to mel-filterbank features [87].

More recently, there has been a shift in the feature extraction stage from a hand-crafted approach to a data-driven strategy, focusing on learning relevant features directly from the raw speech signal. This approach, initially proposed by [88], was motivated by the understanding that hand-crafted features are often the results of analysis of adult speech, and they may not be optimally suited for the acoustic variability present in children's speech. In the study conducted by [88], data-driven feature extraction was performed using CNN-based models. The results indicated that the features learned in a data-driven manner outperformed standard hand-crafted features, emphasising the potential benefits of adapting feature extraction to the specific characteristics of children's speech. Another similar approach was proposed by [89], where the convolutional layers of the feature extractor were replaced by SincNet layers [90]. SincNet uses rectangular band-pass filters instead of the standard CNN filters, enabling a reduction in the number of parameters required for raw waveform modelling. Additionally, this approach involves restricting the filter functions rather than having to learn every tap of each filter.

An alternative approach to mitigate the acoustic variability in children's speech involves working directly with adult features and adapting children's features to reduce the acoustic variability. One commonly used technique for this purpose is Vocal Tract Length Normalisation (VTLN), which has been widely employed to normalise spectral features into a canonical space [91, 92]. Research conducted using VTLN indicates a higher recognition rate when the ASR system is trained with adult speech and subsequently tested with normalised children's speech [93, 94]. Indeed, [95] showed a strong relationship between the optimal warping factor and the age of speakers. VTLN is typically applied as a front-end processing step at the end of the feature extraction. The VTLN process involves stretching or compressing the frequency axis of the spectrum according to a warping function. This process leads to a normalisation of the spectral representation, reducing the impact of speaker variance. Nevertheless, it is important to acknowledge that recognition results achieved with VTLN compensation alone may still be sub-optimal. This is attributable to the presence of various factors, extending beyond variations of the length of the vocal tract, that contribute to the distinct characteristics between adult and children's speech. Extending the VTLN, researchers have explored the normalisation of other specific aspects of children's speech to better align with adult speech characteristics. Several studies investigated the use of pitch normalisation [87, 96–98], while other directly normalised formant values [99, 100]. Furthermore, adapting speaking rate through time-scale modification approaches has been investigated [101]. Children's speech is typically slower than adults, and adjusting the speaking rate can help in creating a more consistent representation for ASR models.

Moreover, in line with the trend of transitioning from knowledge-based to data-driven approaches, some recent studies have explored data-driven feature adaptation methods using deep learning. For instance, studies like [102, 103] employed adversarial multi-task learning to generate age-invariant features.

The goal was to minimise the acoustic mismatch between adult and children's speech by leveraging adversarial training techniques. Adversarial training introduces a form of competition between two neural networks, with one network generating features used by the ASR and another network attempting to distinguish between the adapted children's features and the real adult's features. This adversarial approach aimed to extract features that are less influenced by age-related variations, contributing to improved model generalisation across different age groups.

In addition to feature extraction and feature-level adaptation, certain studies have emphasised the efficacy of appending supplementary information to the acoustic features. For instance, it is a common practice to concatenate speaker embeddings, such as i-vectors [104], to the acoustic features to achieve a more speaker-independent model [105]. Speaker embeddings are compact, fixed-size representations of the features of a speaker's voice derived from their speech signals. Concatenating speaker embeddings with acoustic features enables the model to be more robust to speaker variability, as it incorporates an explicit representation of it. Similarly, both [106, 107] proposed to concatenate various prosodic features, including loudness, voice intensity, and voice probability, with standard acoustic features. This approach has demonstrated success in reducing inter-speaker variances and enhancing discrimination between phoneme classes.

2.3.2 Pronunciation and language model

The conventional information provided in speech corpora typically includes audio signals, corresponding text transcriptions, and anonymised speaker identifiers. However, augmenting this data with additional information holds promise for improving children's speech recognition systems. One pertinent aspect is incorporating the speaker's age, a critical factor, as described in section 2.1, that could facilitate the development of age-dependent ASR systems [108,109]. Moreover, annotating at the sub-word level, rather than the word level, has demonstrated increased performances, particularly in addressing challenges such as mispronunciations or hesitations [110]. This approach enhanced robustness by focusing on smaller linguistic units allowing more flexibility. Another strategy involves the implementation of a children-specific pronunciation model, as illustrated in [111,112]. These lexicons are specifically created to handle pronunciation divergences from canonical adult patterns with knowledge-based children patterns. Finally, the creation of language models explicitly tailored for children's speech can further improve the recognition accuracy [28, 29]. These models capture the linguistic nuances and variations intrinsic to children's language. The integration of these strategies collectively contributes to the development of more effective and adaptive ASR systems for children.

2.3.3 Design of acoustic models

The acoustic model plays a pivotal role in ASR for children's speech, given that the impact of acoustic variability on recognition accuracy degradation is more pronounced compared to linguistic variability. Consequently, the design of the acoustic model is crucial for ensuring robustness to the specific characteristics of children's speech.

Initially, the transition from monophone to triphone HMM-GMM models helped improve performance by capturing co-articulation effects [27, 94]. However, despite a significant portion of children's ASR research relying on HMM-GMM models, as indicated by a literature review [113], acoustic models for children naturally align with the latest advancements in acoustic modelling for adults. Therefore, the design transitions from HMM-GMM to HMM-DNN, as proposed by [6].

The limitations of traditional fully connected neural networks in providing sufficient contextual information prompted their replacement by TDNN layers, as proposed in [114]. TDNNs function similarly to one-dimensional convolutional neural networks. At each time step, both the current time-step frame and its corresponding left and right context are considered, in contrast to traditional DNNs that focus solely on the current time step. To further improve the model's capacity to capture extensive contextual information, multiple layers of TDNN can be used. In addition, motivated by the large overlaps between neighbouring input contexts sub-sampling was introduced, allowing gaps between frames in the context window.

While TDNN was demonstrated as a successful design for children ASR acoustic model [115], they were quickly replaced by Factorised Time-Delay Neural Network (TDNN-F). Indeed, TDNN-F was introduced as an improvement of regular TDNN [116] by decomposing the weight matrix using Singular Value Decomposition (SVD). A more detailed explanation of TDNN and TDNN-F will be provided in section 3.2. These enhancements were especially proven effective for children's ASR [7], outperforming both GMM and TDNN approaches in multiple children datasets. This efficacy can be attributed to the SVD factorisation which divided the weight matrix into two smaller rank matrices, functioning as bottleneck layers. To this day, TDNN-F are the state-of-the-art design for efficient HMM-based children ASR system.

These advancements in neural network architectures, such as TDNN and TDNN-F, showcase the ongoing efforts to refine acoustic models for children's speech recognition, addressing specific challenges and optimising model capabilities to adapt to the unique characteristics of children's speech.

2.3.4 End-to-end models

The success of the end-to-end paradigm in outperforming traditional HMM-based models across various ASR adult datasets has prompted exploration in the domain of children's ASR. However, when attempting to train end-to-end models from scratch using size-limited children's datasets, these models were found to underperform compared to their HMM-based counterparts [117]. To address this challenge, a different training approach was adopted, leveraging pre-trained adult models as a starting point for training, this strategy will be explained in more detail in section 2.3.6.A. The conjunction use of this transfer learning strategy and pre-trained end-to-end models were found to be able to outperform HMM-based models. In their experiments, [117–120] explored various end-to-end architectures as pre-trained models for children's ASR. The architectures investigated included Listen, Attend, and Spell [121], Recurrent Neural Networks (RNNs), ResNet [122], and Transformer [123]. Among these architectures, [117] demonstrated that the Transformer model using a mix of Sequence-to-sequence and CTC losses emerged as the most effective in many cases, demonstrating promising results in the context of children's ASR.

2.3.5 Data augmentation

The success of deep learning can be attributed to its ability to leverage extensive datasets for effective pattern recognition. The depth and complexity of deep neural networks enable them to automatically learn hierarchical representations from data, uncovering intricate patterns and features that may be challenging for traditional machine learning approaches. This adaptability to large datasets contributes to the robustness and generalisation capabilities of deep learning models . However, the scarcity of children's speech data significantly contributes to performance deterioration as compared to adults. This challenge is even more pronounced for languages other than English, where fewer resources are generally available. To address this problem of data scarcity, researchers have explored various data augmentation approaches with the aim of artificially increasing the amount of training data. In the literature, there are two main approaches to data augmentation: using solely the data that is currently available or incorporating external data from diverse sources.

2.3.5.A Using external data

The most natural source of speech data to augment children's speech training data is the abundant reservoir of adult speech data. Studies [124, 125] validate the idea that leveraging out-of-domain adult speech data effectively enhances the automatic recognition of children's speech. Notably, improvements were observed when incorporating adult female speech, given the inherent narrower frequency ranges mismatch between females and children compared to adult male speech. Similarly, the study presented in [126] suggested augmenting speech training data by directly incorporating additional children's data. The observed improvements underscore the significance of leveraging a diverse range of children's speech samples in the augmentation process.

Beyond traditional sources of speech data, researchers have delved into the use of synthetic data as a supplementary resource for training children's ASR. The idea behind employing synthetic speech data revolves around generating speech signals that perceptually resemble a child. In this regard, voice conversion has emerged as a notable method. Voice conversion involves leveraging extensive adult datasets and transforming them into children's speech while preserving the content. Various voice conversion approaches have been applied to the generation of synthetic children's speech, encompassing classical signal processing manipulations such as vowel stretching [127], fundamental frequency shift [128], and spectral envelope wrapping [129]. Additionally, studies like Shuyang's work [130] have investigated the combined use of these signal processing manipulations to further improve the ASR system. In contrast to signal processing modifications, deep learning methods, particularly Generative Adversarial Networks (GANs), were found effective for voice conversion. In [131], a GAN was used to transform children's speech into adult-like speech. Therefore, this approach can use regular adult speech as data augmentation rather than converting adult speech into children's speech. This strategy aims to reduce variability by training the generator to produce adult-like speech directly from children's speech. The GAN model comprises a generator responsible for creating synthetic data and a discriminator distinguishing between generated and real data samples in an adversarial way. During inference, the discriminator is removed, leaving only the generator to convert children's speech into adult-like speech.

Besides voice conversion, Text-to-Speech (TTS) systems have been using to generate speech examples directly from text. Recent advancements in TTS systems, such as Tacotron2 [132] and Variational Inference with adversarial learning for end-to-end Text-to-Speech (VITS) [133], have enhanced the realism of generated speech utterances. Consequently, some researchers have explored the use of TTS outputs as data augmentation for adult ASR tasks [134]. However, children's speech exhibits more complex traits than adults, including substandard or unclear pronunciation and acoustic variability. As a result, the quality of children's TTS is often inconsistent. To address this, [135] proposed data selection strategies based on speaker embedding similarity between the reference speaker and the speaker embedding extracted from generated speech utterances. Hence, eliminating synthetic utterances that significantly deviate from their reference examples. This approach significantly improved the recognition score for various children's speech recognition tasks.

2.3.5.B Using available data

In scenarios where the inclusion of external data is not possible, there is a necessity to enhance model robustness by directly modifying the existing dataset. An established approach to enhance the model robustness involves generating augmented versions of the original data, by adding diverse acoustic perturbations to them. Typically, these perturbations are additive noise, babel noise, white noise, music and reverberation [136–140]. This augmentation strategy which introduces noise and reverberation into the existing dataset, aims to simulate real-world conditions where environmental factors can impact the recognition of speech signals.

An alternative strategy for augmenting the original speech dataset involves creating copies where the dimensions of the speech signal are perturbed. These perturbations include modifications along the time axis, as demonstrated by speed perturbation to better match children's speaking rate variability [141], and modifications along the frequency axis [142] through vocal tract length perturbation [143], simulating variations in vocal tract dimensions. More recently, techniques like SpecAugment [144] were found particularly effective, especially in the end-to-end paradigm. SpecAugment involves random masking of frequencies and time bands within the spectrogram in conjunction with time and frequency warping.

Finally, as mentioned in Section 2.1.2, it is crucial to recognise the presence of disfluencies and errors in children's speech, presenting inherent challenges to the learning process of ASR models, especially in reading tasks. Therefore, to enhance the model's robustness in handling such errors, a noteworthy proposition by [145] involves the manual creation of synthetic reading errors. This involves manual interventions, specifically achieved by cutting the signal, resulting in a deletion error, or incorporating speech elements produced by other children to simulate substitution or insertion errors.

2.3.6 Training procedure for children speech recognition

In the domain of HMM-based ASR systems for children, several strategies have been employed to adapt acoustic models for enhanced performance. Notably, adaptation techniques such as Maximum Likelihood Linear Regression (MLLR) and Maximum A-Posteriori (MAP) have been applied with success. Additionally, the use of Speaker Adaptive Training (SAT), specifically based on Feature MLLR (fMLLR) or Constrained MLLR (CMLLR), has been proven effective at improving the performances of ASR systems designed for children [1, 29, 111, 146].

As mentioned in Section 2.1.3, the integration of DNNs into children's ASR systems necessitates a substantial amount of labelled data to provide optimised performances. The efficacy of DNNs rely on their two-pass iterative training procedure, involving a forward and a backward pass. In the forward pass, the input corpus is fed to the network, generating prediction outputs. Subsequently, the loss is computed by comparing these predictions with the ground truth target values. In the backpropagation phase, the objective is to mitigate prediction errors by adjusting the network weights using the gradient descent technique and leveraging the computed loss. The training process continues through multiple iterations of these forward and backward passes until the model converges.

In response to the challenges posed by the distinctive variations in children's speech, novel adaptations of this training pipeline have been proposed. Noteworthy among these is transfer learning, which leverages efficient pre-trained DNN models. Additionally, multi-task learning has been explored to capture shared representations across related tasks, while self-supervised learning has emerged as a promising paradigm that enables the model to learn information about speech without relying on labelled data.

2.3.6.A Transfer learning

When confronted with a new problem, humans exhibit the ability to draw upon information from prior tasks as an inductive bias. This cognitive capacity enables individuals to avoid starting the learning process entirely from scratch by leveraging knowledge acquired from past tasks. This ability, often referred to as Transfer Learning (TL), may be defined as the capacity to identify and use knowledge from previous tasks as a foundation for approaching new tasks.

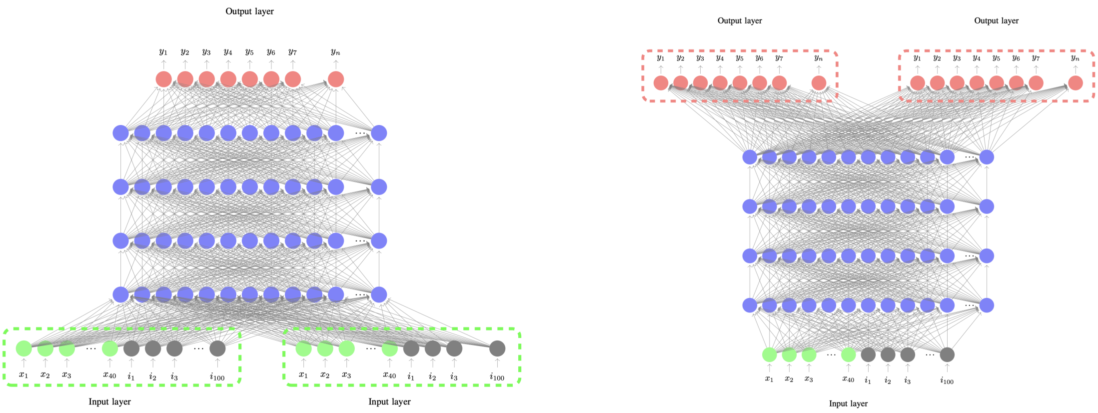
In contrast, in the context of machine learning, algorithms are generally developed from scratch on a specific task, lacking the inherent capability to transfer knowledge. Therefore, the concept of TL, or parameter transfer, has emerged as a pivotal bridge between artificial and biological intelligence. In this paradigm, a model's parameters are initialised with values derived from another well-resourced model trained on a related source task. Subsequently, the model's parameters are adapted, also called fine-tuned, with data from the new domain, adjusting parameters to better align with the target task. This knowledge transfer allows the model to leverage underlying characteristics acquired from the source task, contributing to enhanced performance when applied to the target task.

Furthermore, a notable advantage of TL is its ability to require a reduced amount of training data for adaptation. Building upon a pre-trained model, it leverages the knowledge acquired during the initial training on a source task. The target model can exploit the information encapsulated in the pre-trained parameters, mitigating the demand for an extensive target domain dataset. The reduced need for adaptation training data makes transfer learning particularly advantageous in scenarios where labelled data is scarce or challenging to obtain, such as in children's speech.

A common assumption in deep learning TL is that the lower layers, situated closer to the input, tend to capture more signal-specific characteristics, whereas the higher layers, in proximity to the output, capture more task-specific information [147, 148]. This hierarchical organisation within neural networks aligns with the notion that lower layers are learning general representations of input data, or low-level features, while higher layers specialise in extracting intricate patterns that are specifically relevant to the task at hand. Therefore, depending on the target task, adaptation may be more beneficial in either the top or lower layers of the model.

In recent years,TL has emerged as a highly successful technique across various applications, particularly proving effective in low-resource tasks such as language understanding [62], character recognition [149], and dysarthric speech recognition [150], among others. The achievements in these domains have spurred interest in exploring the utility of transfer learning for children's speech recognition, a domain often characterised by limited labelled data.

Given the prevalence of large corpora of adult speech, recent acoustic models trained on adult data have demonstrated high efficiency and encapsulate rich acoustic and phonetic information. Motivated by these successes, [6] proposed investigating three distinct transfer learning methods to assess the contribu-



(a) Acoustic adaptation

(b) Pronunciation adaptation

Figure 2.10: Transfer learning approaches. Figures from [6]

tions of acoustic adaptation, pronunciation adaptation, and their combination in the context of children's speech recognition.

Acoustic adaptation targets the lower-level layers, leveraging the established notion that these layers capture acoustic properties. The methodology involves freezing the weights of the top-level layer and applying transfer learning to the lower-level layers, as depicted in Figure 2.10(a). In experiments conducted by [6] and [151], acoustic transfer learning from an adult model to children's speech, by retraining only the first layers, yielded substantial relative WER improvements of 38% and 26%, respectively, compared to the performance of adult models. Impressively, the acoustic adaptation outperformed a randomly initialised acoustic model trained on the same children's data, achieving a 4.9% relative WER improvement.

Pronunciation adaptation, premised on the idea that higher-level layers capture task-specific information, focuses on adapting these layers while keeping lower-level layers frozen. As illustrated in Figure 2.10(b), [6] conducted experiments applying pronunciation adaptation to the last layers of the model, resulting in a significant 31% relative WER improvement compared to the performance of the adult model. However, when compared to a randomly initialised acoustic model trained with the same children's data, pronunciation adaptation exhibited a 5.6% relative WER degradation.

Finally, the combination of acoustic and pronunciation adaptation, achieved through fine-tuning the entire network, demonstrated outperforms performances compared to individual adaptations. This aligns with recent observations in end-to-end models, where transfer learning from an adult pre-trained model outperformed training from scratch using only children's data [117, 118]. These findings underscore the efficacy of transfer learning strategies for optimising acoustic and pronunciation adaptation in children's ASR.

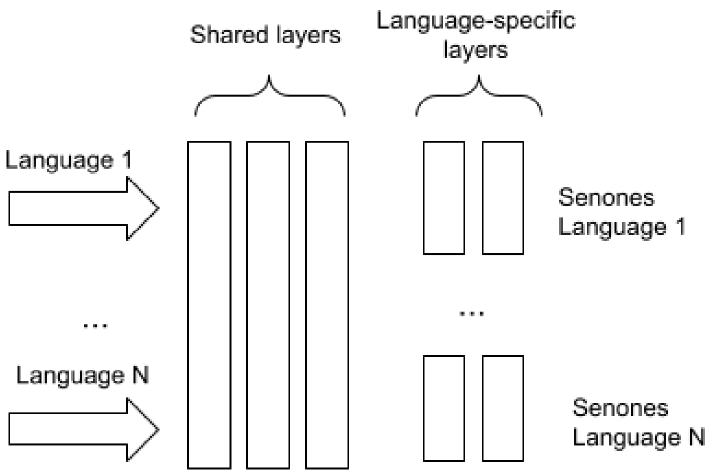


Figure 2.11: Multilingual approach using each language as a task in a multi-task learning context.

2.3.6.B Multi-task learning

Multi-task learning (MTL), like to transfer learning, draws inspiration from biological intelligence. In contrast to transfer learning, MTL does not train solely on source and target tasks in a sequential manner, here MTL simultaneously train on multiple tasks at the same time. The fundamental objective of MTL is to discover shared representations among related tasks. In general, a typical MTL model consists of two distinct components. The first part is a sub-network shared by all tasks, while the second part consists of task-specific output sub-networks, as illustrated in Figure 2.11. The shared layers facilitate the learning of a joint representation that is more robust, enhancing the model's reliability across diverse tasks.

More formally, for any task i the corresponding output of the forward pass will be:

$$f(X_i; \{M_i, M_c\}) = f_i(f_c(X_i, M_c); M_i) \tag{2.12}$$

where X_i is the data associated with the task i, M_i represents the task-specific parameters of the model, and M_c corresponds to the parameters that are shared (or common) across all tasks.

In consequence, the performances of MTL are intricately tied to the degree of relatedness among tasks used during training. Indeed, its efficacy diminishes when confronted with outlier tasks that are unrelated to the majority of the other tasks. This sensitivity is due to the inherent challenge of learning common representations for tasks that lack substantial relatedness to one another [152]. This task-relatedness consideration underscores the importance of thoughtful task selection when using MTL techniques.

Moreover, MTL has been used effectively in a variety of areas, including natural language processing [153], computer vision [154] and bioinformatics [155]. Naturally, MTL has been applied in the field of automatic speech recognition [156] with direct application to low-resource ASR [157]. Given that ASR for children represents a resource-limited task, MTL has been proposed as a strategy to mitigate the issue of data scarcity. Notably, studies such as [151] and [158] successfully applied MTL to Mandarin and English-speaking children, with a 16.96% relative improvement in WER for the English children.

2.3.6.C Self-supervised Learning

A first step towards Self-Supervised Learning (SSL) was the introduction of semi-supervised techniques, such as pseudo-labeling. Semi-supervised approaches were the dominant training strategies for using unlabeled data. In particular, the pseudo-labeling starts with the training of a "teacher" model on a set of supervised data. Subsequently, pseudo-labels are generated for unlabeled data by leveraging the predictions of the trained teacher model. Following this, a "student" model is trained using a combined dataset comprising both supervised and pseudo-labeled data. Importantly, the pseudo-labeling process can be iteratively repeated multiple times to enhance the quality of teacher-generated labels [159, 160]. It is noteworthy that, as of today, some of the most effective ASR models, such as the Whisper model, leverage pseudo-labeling as a crucial component of its training strategy [30]. However, the performances of semi-supervised, and particularly pseudo-labelling, have been found to be highly dependent on the quality of the teacher model, prompting the need for more robust and sophisticated training strategies.

In this context, SSL has emerged as a paradigm designed to acquire general data representations directly from unlabeled examples, subsequently allowing transfer learning on a small amount of labelled data. This approach has proven particularly successful in the domain of natural language processing [161] and computer vision [162]. One notable first attempt to bring SSL to the speech domain was made by the introduction of the Problem-Agnostic Speech Encoder (PASE) and its extension,PASE+. These innovations demonstrated the capability to learn meaningful speech information such as speaker identities, phonemes, and emotions. The PASE framework operates by encoding raw speech waveforms into a learned representation, which is then input to multiple regressors and discriminators. The regressors within PASE are standard features computed from the input waveform. While the discriminators focus on positive or negative samples and are trained to effectively separate them. Both the regressors and discriminators play a crucial role in incorporating prior knowledge into the encoder, a key factor for deriving meaningful and robust representations.

Recently, SSL systems have obtained remarkable results with the introduction of models employing BERT-like training methodologies, such as Wav2vec2 [163] and HuBERT [164]. Notably, the success of these models can be attributed to the conjunction use of masking, discrete speech units, contextualised representations, and contrastive loss. The integration of masking techniques allows these models to effectively learn contextualised representations by masking certain portions of the input data and predicting them based on the remaining context. The use of discrete speech units enables the model to be more robust to variations, while the contrastive loss functions enhance the discriminative power of these models by encouraging the model to differentiate between positive and negative samples.

Motivated by the capabilities of SSL methods in overcoming challenges in low-resource ASR tasks, such as low-resource languages [165], noisy speech [166], and accented speech [167], the integration of SSL for children's ASR marked its debut in 2021, with a first place in a non-native children's speech recognition challenge [168]. Subsequent to this notable success, the application of SSL for children's ASR has gained increased attention, especially with the use of models like Wav2vec2 [169–171]. A concise analysis of various SSL approaches as frozen feature extractors for children's ASR has been conducted within the context of this thesis, and the findings are presented in Annex B.

2.4 Children Corpora

As described in Chapter 2.1.3, notwithstanding recent efforts to assemble dedicated databases for children's speech, the quantity of available data remains lower compared to adults. Collecting speech data from children is challenging in many ways, from factors such as limited attention spans, frequent mispronunciations, ungrammatical expressions, and the use of non-standard vocabulary. These difficulties involved in capturing high-quality child speech data contribute to the scarcity of publicly accessible child speech corpora. Additionally, the relatively modest sizes of these datasets present obstacles to research efforts and impede progress in developing reliable ASR systems for children.

Table 2.1 provides a compilation of existing corpora of children's speech. Notably, approximately one-third of the available corpora are in English. Likewise, a comparable proportion is specifically oriented towards children under the age of 4, where the speech dataset comprises child-adult interactions (usually with the parents). It is crucial to acknowledge the inherent trade-offs in these different corpora, involving considerations of speaker diversity, total duration, and the number of utterances.

Subsequently, the remainder of this section will provide a more detailed description of the children's speech corpora employed in this thesis.

2.4.1 LETSREAD

LetsRead database [191] is a read-aloud speech database of European Portuguese from children aged 6 to 10, from 1st to 4th grade. This corpus is composed of a total of 284 children, 147 girls and 137 boys, whose mother tongue is European Portuguese. Children from private and public Portuguese schools were asked to carry out two tasks: reading sentences and a list of pseudo-words. The difficulty of the tasks varies depending on the school year of the child. For this proposal, we excluded all utterances from the pseudo-word reading task because we do not include pseudo-words in the language model and lexicon in our experiments.

Corpus	Languages	# Spkrs	# Utt	Dur.	Age Range	Date
Providence Corpus [172]	English	6		363h	1-3	2006
Lyon Corpus [173]	French	4		185h	1-3	2007
CASS_CHILD [174]	Mandarin	23		631h	1-4	2012
Demuth Sesotho Corpus [175]	Sesotho	4	13250	98h	2-4	1992
NITK Kids' Speech Corpus [176]	Kannada	160		10h	2-6	2019
CHIEDE [177]	Spanish	59	15,444	8h	3-6	2008
CUChild [178]	Cantonese	1,986			3-6	2020
EmoChildRu [179]	Russian	100	20,000	30h	3-7	2015
CNG Portuguese children [180]	Portuguese	510		21h	3-10	2013
AusKidTalk[1] [35]	English	750		600h	3-12	2021
UCLA JIBO kids [181]	English	130			4-7	2019
PF-STAR-SWEDISH [182]	Swedish	198	8,909	6h	4-8	2005
SLT 2021 [183]	Mandarin	981		58h	4-11	2021
PF-STAR Children British [38, 182, 184]	English	158		14.5h	4-14	2006
AD-child. RU [185]	Russian	278			4-16	2019
TBALL [186]	English	256	5,000	40h	5-8	2005
SPECO [187]	Hungarian	72		12h	5-11	1999
UltraSuite [188]	English	86	14,456	37h	5-14	2019
CID read speech corpus [189]	English	436			5-18	1996
Persian Kids Speech Corpus [190]	Persian	286	162,395	33h	6-9	2022
Letsread[2] [191]	Portuguese	284	4,629	14h	6-10	2016
CMU kids Corpus [36]	English	76	5,180		6-11	1997
CFSC [192]	Filipino	57		8h	6-11	2012
IESC-Child [193]	Spanish	174	19,793	34h	6-11	2020
CU Children's read and prompted [194]	English	663	66300		K-G5	2001
Chorec[2] [195]	Dutch	400	3,065	25h	6-12	2008
ChildIt2 [196]	Italian	96	4,875	9h	6-14	2016
TIDIGITS [197]	English	101			6-15	1993
CSLU Kids' Speech Corpus [37]	English	1,100	1,017		K-G10	2007
SingaKids-Mandarin [34]	Mandarin	255	79,843	125h	7-12	2016
ChildIt corpus [198]	Italian	171			7-13	2007
VoiceClass Database [199]	German	170			7-14	2010
Deutsche Telekom telephone [199]	German	106			7-14	2010
Jasmin [200]	Dutch			63h	7-16	2008
Tgr-child corpus [198]	Italian	30			8-12	2007
SponIt corpus [198]	Italian	21			8-12	2007
Swedish NICE Corpus [201]	Swedish	75	5,580		8-15	2005
CHIMP spontaneous speech [27]	English	160			8-14	2002
SpeeCon corpus [202]	20 Languages				8-15	2002
Rafael.0 telephone corpus [108]	Danish	306			8-18	1996
Boulder Learning - MyST [33]	English	1,371	228,874	384h	G3-G5	2019
CU Story Corpus [194]	English	106	5,000	40h	G3-G5	2003
ETLT[2] [203]	L2 German		1,674	6h	9-16	2020
Lesetest corpus [204]	German	62			10-12	2000
FAU Aibo Emotion Corpus [205]	German	51	13,642	9h	10-13	2002
PIXIE corpus [206]	Swedish	2,885				2003
Takemaru-kun corpus [207]	Japanese	17,392				2007
CALL-SLT [208]	German		5,000			2014

[1] To this day, data collection for this dataset is not complete.
[2] Information displayed here correspond to a subset of the original data used in this proposal.

Table 2.1: Non-exhaustive comparison of children's speech corpora. This table has been sorted by age range. Blanks indicate unavailable information. Entries highlighted in bold correspond to the corpora used in the experiments presented in this thesis. K: Kindergarden. G: Grade

2.4.2 PFSTAR_SWEDISH

The PFStar children's speech corpus [182] was collected as part of the EU FP5 PFSTAR project. It contains more than 60 hours of speech. This corpus is divided into two parts: native-language speech and non-native language part. The native-language speech part contains recordings of British English, German and Swedish children, from 4 to 14 years old. The non-native language part consists of speech by Italian, German and Swedish children speaking English. In this work, we only used the native language Swedish part, consisting of speech by 198 native Swedish children, between 4 and 8 years old recorded in the Stockholm area, imitating an adult who read the text from a screen.

2.4.3 ETLTDE

Extended Trentino Language Testing (ETLT) corpus [203] has been collected in northern Italy for assessing English and German proficiency of Italian children between 9 and 16 years old, by asking them to answer questions. The data collection was carried out in schools. On average the signal quality is good, but some background noise is often present (doors, steps, keyboard typing, background voices, street noises if the windows are open, etc). In addition, many answers are whispered and difficult to understand. For this thesis, we only used the German-transcribed subset, named ETLTDE, a subset containing around 6 hours of speech divided into training and test partitions.

2.4.4 CMU_KIDS

The CMU kids corpus [36] contains English sentences read aloud by children, 24 males and 52 females, from 6 to 11 years old. In total, 5,180 utterances were recorded with one sentence per utterance. This database was created to train the SPHINX II [209] automatic speech recognition system within the LISTEN project at CMU.

2.4.5 CHOREC

The Chorec corpus [195] consists of 400 Dutch-speaking elementary school children, between 6 and 12 years old, reading words, pseudo-words and stories. The difficulty of the reading task was adapted to children with 9 different levels. Recordings were made in schools, leading to some environmental noises (school bells, children entering the playground etc.). For this thesis, similarly to the LETSREAD dataset, we discarded pseudo-word utterances.

2.4.6 MyST

My Science Tutor (MyST) Children Speech Corpus [33] is currently one of the largest publicly available corpora of English children's speech, with around 400 hours. This is about 10 times more than all other English children's speech corpora combined. It consists of conversations between children and a virtual tutor in 8 scientific domains. Speech was collected from 1,372 students in the third, fourth and fifth grades. Partitioning of the corpus is already available, ensuring a reasonable representation of each scientific domain and that each student is present in only one partition. However, only 45% of the utterances were transcribed at the word level. Furthermore, for the purposes of all our experiments in the thesis, we decided to remove all utterances shorter than one second and longer than 20 seconds and shorter than one second. Typically, utterances shorter than one second were found to predominantly contain silence alone, while those longer than 20 seconds were constrained by our GPU limitations. After this filtering, 81971 utterances from 736 speakers for a total of 151 hours remain.

2.5 Summary

In this chapter, we provided an overview of children's ASR, its inherent challenges, and the ongoing responses from the research community. The complexity of ASR in children arises primarily from the developmental nuances of the vocal apparatus, resulting in an acoustic mismatch with adult speech, although this mismatch gradually diminishes until around the age of 15. Despite sustained efforts, ASR for children remains an active and challenging area of research.

Our examination in this chapter highlights that knowledge transfer approaches, such as transfer learning and multi-task learning, appear to be promising avenues for improving children's ASR. Additionally, the exploration of synthetic speech generation, using TTS systems, has captured our attention as a potential strategy for improvement. These methodologies will be applied and explored in-depth in the subsequent chapters of this thesis.

Moreover, we have presented a comprehensive comparison of children's speech corpora, which, to the best of our knowledge, stands as the most exhaustive compilation available. This analysis provides valuable insights into the landscape of available datasets for children's speech.

3

Hybrid models for children automatic speech recognition

Contents

3.1 Introduction

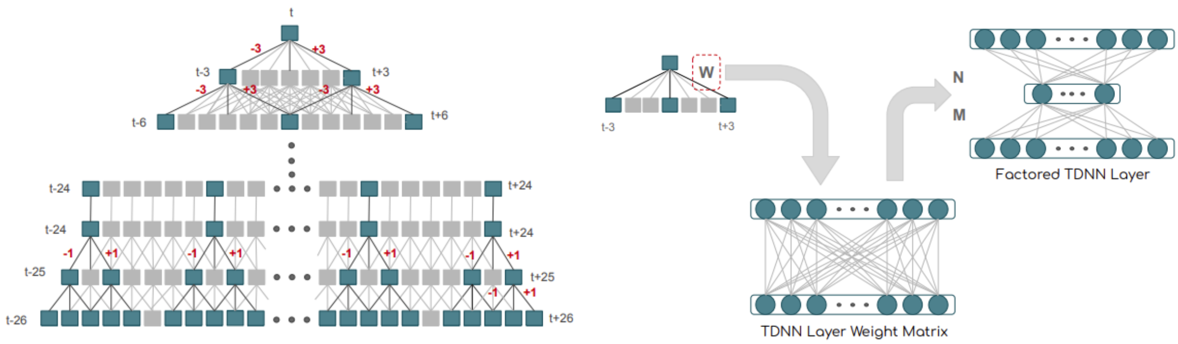
In the history of ASR, HMM-based models emerged as a popular choice for acoustic modelling since the 1980s. The application of HMMs offered a structured framework to effectively model the temporal dependencies inherent in speech signals. Initially, the paradigm involved combining HMMs with GMMs to represent the probability distributions of acoustic features. This traditional HMM-GMM architecture, while effective, faced limitations in capturing the intricacies of complex, non-linear relationships present in speech data. Therefore, a pivotal turning point occurred with the integration of DNN, evolving towards hybrid HMM-DNN models. This paradigm shift led to substantial improvements in ASR performances. Indeed, DNNs brought enhanced modelling capabilities, allowing the system to discern more nuanced acoustic patterns and adapt better to diverse linguistic contexts. The hybrid HMM-DNN approach has become key architecture in modern ASR systems, demonstrating remarkable success in handling large vocabulary tasks and challenging acoustic conditions [46].

In Chapter 2, we explained the integration of acoustic, phonetic, and linguistic knowledge within the ASR pipeline, particularly in the context of HMM-GMM and HMM-DNN systems. This integration includes the cooperation of multiple modules, such as the vocabulary, acoustic model, and language model, which all work together to support the Decoder's operations. By directly integrating hand-crafted knowledge into the ASR system, the quantity of speech data needed to produce suitable results can be decreased. These characteristics were proven to be beneficial for children ASR where the amount of data is often limited. For years, HMM-based configurations have been a prominent choice in children's ASR research. Between 2009 and 2020, 80% of published research on children's speech recognition was based on HMM systems, with 45% using HMM-GMM and 35% HMM-DNN. The literature suggests that ASR results can be improved through the application of inductive bias approaches such as transfer learning and multi-task learning [151]. However, it was noted that during the same period, 63% of published work focused on English [210]. Therefore, it remains uncertain how children's speech from other languages relates to the various approaches used in English, particularly transfer and multi-task learning.

In this chapter, we will first present the TDNN-F architecture used for the development of our HMM-DNN ASR system. Then, we will delve into applying knowledge transfer strategies across various scenarios, encompassing both non-English and English datasets. Applying it to the non-English dataset will allow us to understand if inductive bias approaches are still working in the context of non-English children's speech. To this end, we present an examination of both transfer and multi-task learning methodologies, leveraging adult speech as an inductive bias, a prevalent approach in the existing literature.

Finally, as part of our contribution, we introduce a novel approach called "multilingual transfer learning." This strategy integrates both transfer and multi-task learning techniques. Here, we address the challenge posed by the scarcity of data for both adults and children in low-resource languages by employing a combination of transfer and multi-task learning methodologies solely on low-resourced children's speech datasets.

3.2 Factorised Time Delay Neural Network for children ASR



(a) TDNN with sub-sampling **(b)** Factorised TDNN layer

Figure 3.1: TDDN and TDNN-F taken from [7]

In this chapter, we will focus on the application of HMM-DNN-based models, specifically leveraging the TDNN-F architecture. Instead of revisiting the entire HMM-DNN framework, extensively covered in Chapter 2, we will delve into the distinctive features of the DNN architecture employed in our HMM-DNN ASR systems, namely the TDNN-F. The use of the particular TDNN-F architecture instead of DNNs or TDNN was motivated by the need to improve ASR performances in scenarios characterised by limited speech data and its successful results for children ASR [7]. This choice aligns with the broader trend in ASR research where there is increasing interest in exploring more data-efficient neural network architectures, with the TDNN-F standing out as a notable example.

In a TDNN layer, the input at time step t is sent to the next layer along with its neighbouring frames within a specific context window [114]. The multiple TDNN layers enable the integration of information over a broader context window. However, this design can lead to computational inefficiency due to the overlap of adjacent steps context windows. In response to this computational issue, an alternative approach involves sub-sampling the input sequence, skipping certain frames and transmitting only the sampled frames. This sub-sampling process is illustrated in Figure 3.1(a). The sub-sampling technique offers a trade-off between computational efficiency and the network's ability to capture temporal dependencies. This is based on the assumption that neighbouring activations are correlated. Therefore, by selectively processing a subset of frames, the TDNN can maintain its ability to capture relevant information while reducing the computational cost associated with processing every frame.

The TDNN-F architecture [116] enhances the TDNN by improving the computational efficiency of the network. This is achieved by decomposing the weight matrix of each TDNN layer into an approximation as the product of two lower-rank matrices using SVD, defined as:

$$\mathrm{W} = \mathrm{U}\Sigma\mathrm{V}^T = \mathrm{MN} \tag{3.1}$$

Here, $\Sigma \in \mathbb{R}^{m \times n}$ is a non-negative rectangular diagonal matrix, and $\mathrm{M} \in \mathbb{R}^{m \times k}$ and $\mathrm{N} \in \mathbb{R}^{k \times n}$ with $k \leq \min(m, n)$. The goal is to ensure that one of the two sub-matrices is close to a semi-orthogonal matrix, representing either $\mathrm{U}\Sigma$ or $\Sigma\mathrm{V}^T$. Notably, k is typically chosen to be much smaller than m and n. During the training of the network, after every few updates of the entire network, a specific update is performed on matrix N using Stochastic Gradient Descent (SGD). This update is guided by an additional objective function, which ensures that N is not too far from being semi-orthogonal. The introduction of the semi-orthogonal constraint through the matrix factorisation process in TDNN-F can be viewed as adding an extra “bottleneck layer” to the traditional TDNN. A visual representation of the TDNN-F layers is presented in Figure 3.1(b). This “bottleneck layer”, represented by the reduced-rank matrix N, contributes to the overall efficiency of the network by reducing the number of parameters and computations required while preserving essential information.

3.3 Assessing the efficacy of multi-task and transfer learning from adult to children

3.3.1 Methodology

Using adult data as an inductive bias for knowledge transfer approaches in children’s ASR has been found to be beneficial in tackling the challenges specific to children’s ASR, as highlighted in the literature. [6, 151, 158]. The motivation behind leveraging adult data for pre-training lies in the stability and reduced variability found in adult speech. This characteristic simplifies the extraction and recognition of intrinsic and meaningful speech patterns, making it an efficient approach for knowledge transfer to improve children’s ASR. In this initial experiment, our aim is to first validate these findings and subsequently extend these methods to a low-resource language. To this end, we propose to assess children’s speech recognition performances in four distinct configurations:

1. **Adult model**: Training a model from scratch using only adult data. This configuration serves as a baseline to assess the standalone performance of a model trained exclusively on adult speech.

2. **Children model**: Training a model from scratch using only children’s data. This configuration provides insights into the model’s ability to learn from children’s speech without leveraging adult data.

3. **Multi-task model**: Training a model concurrently on adult and children data using multi-task learning. This configuration explores the potential benefits of simultaneous training on both adult and children data.

4. **Transfer learning**: Fine-tuning a model on children's data that was pre-trained on adult data (from the Adult Model in Configuration 1). This configuration assesses the effectiveness of transferring knowledge from adult to children data for improved speech recognition.

By systematically comparing these configurations, our objective is to identify the most effective approach for enhancing performance in children ASR, particularly in low-resource language scenarios and answering the research question *Which knowledge transfer approach is best for efficiently modelling and improving automatic recognition of children's speech?*

3.3.2 Corpus

For our evaluation of knowledge transfer performances in children's speech, we considered both English and a low-resource language, specifically European Portuguese. European Portuguese can be qualified as a low-resource language due to the limited availability of large-scale adult speech corpora, with most datasets not exceeding 100 hours [211]. In our experiments, we utilised a subset of the LibriSpeech corpus for English and the BD-PUBLICO corpus for Portuguese as the adult corpora. For the children's dataset, we employed Myst and LetsRead for English and Portuguese, respectively. In this section, we offer a description of the adult corpora. Detailed information on both children datasets can be found in Section 2.4. The statistics of all corpora used in this initial experiment are presented in Table 3.1.

Language	Corpus name	Train	Test
English	LibriSpeech *Adult*	104014 utterances 2097 speakers 363h	2620 utterances 87 speakers 5h
	Myst *Children*	60897 utterances 566 speakers 113 hours	4079 utterances 91 speakers 13 hours
Portuguese	BD-PUBLICO *Adult*	8085 utterances 100 speakers 22 hours	412 utterances 10 speakers 1 hour
	LetRead *Children*	3590 utterances 180 speakers 12 hours	1039 utterances 52 speakers 2 hours

Table 3.1: Number of utterances, number of speakers, and the duration of training and testing sets for both English and Portuguese corpora, encompassing both adult and children training and test sets

LibriSpeech

The LibriSpeech dataset [31] is a widely used English corpus in the ASR field, introduced to address the need for large-scale, high-quality speech datasets to advance ASR performances. The dataset consists of diverse audio recordings derived from audiobooks obtained from the LibriVox project, totalling approximately 1000 hours of labelled audio sampled at 16 kHz. LibriVox, a community-driven initiative, involves volunteers reading and recording around 8,000 public domain books at the time of the creation of Librispeech. Thus, LibriSpeech encapsulates the natural variability present in real-world spoken language, featuring speakers from diverse backgrounds and reading styles.

LibriSpeech is organised into different subsets, with the training data split into three partitions of 100 hours, 360 hours, and 500 hours. The development and test data are categorised as "clean" and "other" respectively, based on the difficulty levels for ASR systems. In our experiments, we used the 360-hour training set, as it already represents well-resourced scenario conditions.

BD-PUBLICO

The BD-PUBLICO database (Base de Dados em Português eUropeu, vocaBulário Largo, Independente do orador e fala COntínua) [212] consists of reading sentences extracted from the Portuguese newspaper PÚBLICO over a 6-month period, the corpus encompasses linguistic diversity with 10 million words and 156,000 different forms. Recordings involve 120 speakers, specifically graduate and undergraduate students from Instituto Superior Técnico (Lisbon), all falling within the age range of 19 to 28 years, establishing BD-PUBLICO as an adult dataset. Recordings occurred in optimal noise conditions within a soundproof room at INESC-ID (Lisbon), with a sampling frequency of 16kHz and the use of a high-quality microphone. The corpus includes a pronunciation lexicon with citation phonemic transcriptions, with manual corrections applied to enhance transcription accuracy.

For effective model training and evaluation, we partitioned the BD-PUBLICO corpus into three distinct sets, ensuring balanced gender distribution. The training set comprises 80 sentences performed by each of the 100 speakers. The development set, consisting of 40 sentences performed by 10 speakers. Finally, the test set includes 40 sentences performed by 10 speakers.

3.3.3 Experimental setup

All experiments were conducted using the Kaldi open-source toolkit [69]. Initially, for each corpus, an independent HMM-GMM acoustic model was trained to generate the necessary alignment for the training of the HMM-DNN model. Subsequently, HMM-DNN acoustic models were trained using 40-dimensional fbanks alongside 40-dimensional Spectral Subband Centroid (SSC) features [213]. The incorporation of SSC features, which share properties with formant frequencies, is expected to enhance vowel recognition and contribute to improved recognition of children's speech. The resulting 80-dimensional input features were augmented by a 100-dimensional speaker embedding i-vector. Concatenating speaker embeddings to the input features is a commonly employed strategy to enhance model speaker robustness [104]. Our i-vector extractor was specifically trained on a pooled set of children's data, encompassing LetsRead, PfStar Swedish, Etltde, Cmu kids, and Chorec, all described in Section 2.4.

To augment the training corpora, data augmentation techniques were applied, including perturbing the speaking rate of each training utterance by factors of 0.9 and 1.1, as well as volume perturbation. This augmentation strategy enhances the network's robustness to variations in speaking rate and volume during testing. Additionally, Specaugment [144] was employed on top of the fbanks and SSC features, involving random masking of time and frequency bands to further improve model robustness.

For all experiments, a consistent HMM-DNN acoustic model architecture was employed, trained using the Lattice-Free Maximum Mutual Information (LF-MMI) objective function in conjunction with a cross-entropy loss. The learning rate used for training was $2 \cdot 10^{-4}$. The acoustic model architecture is divided into two parts: i) six convolutional neural network layers and seven TDNN-F layers with a dimension of 1024, followed by ii) two TDNN layers with a dimension of 450 and a single fully-connected layer working as an output layer. In the transfer learning experiments, only the first part of the network was fine-tuned, while the second part was replaced by randomly initialised layers. Similarly, in MTL experiments, the first part was shared between adult and child models, while the second part remained independent.

3.3.4 Results

	English		*Portuguese*	
Method	Adult WER ↓	Children WER ↓	Adult WER ↓	Children WER ↓
Adult model	**6.53%**	43.87%	**3.82%**	102.83%
Children model	15.78%	25.53%	45.56%	25.88%
Multi-task model	6.74%	30.56%	4.59%	27.65%
Transfer learning	-	**21.53%**	-	**25.36%**

Table 3.2: WER results using adult data for knowledge transfer methods

The WER scores for all settings are presented in Table 3.2. In the first row, we observed that employing a model trained solely on adult data yields a WER of 43.87% and 102.83% on the children's corpora for English and Portuguese, respectively. Meanwhile, the same adult model achieves WER scores of 6.53% and 3.82% for Librispeech and BD-PUBLICO, respectively. The notable degradation in children's scores compared to adults demonstrates the considerable variability present in children's speech, which has a detrimental impact on the ASR scores. We observed that the Portuguese children's score degradation is higher than the English one, which can be explained by the higher mismatch between the adult and children recording setting coupled with the youngest age range in the Portuguese dataset. This supports the idea that an acoustic model designed exclusively for children is necessary because child speech is currently unusable with adult systems.

On the other hand, training the acoustic model directly on children's data yields substantial improvements of the WER on the children's test sets, reaching 25.53% and 26.88% for English and Portuguese, respectively. This indicates that exposing the model to children's acoustic variabilities during training enhances its robustness to such variations. However, this improvement comes at the cost of deteriorated adult speech recognition performance, resulting in WER of 15.78% and 45.56% for English and Portuguese corpora, respectively. This further confirms the presence of acoustic mismatch between adult and children's speech. Subsequently, we compare the transfer and MTL approaches using these two models trained from scratch as a baseline.

In the MTL scenario, where the model is trained jointly using both adult and children data at the same time, recognition scores for adults and children marginally decrease compared to their individual baseline counterparts. However, unlike the adult and children baseline models, where the acoustic mismatch significantly impacted the WER scores on the mismatch domain, the MTL scenario shows comparable recognition scores to their respective "trained from scratch" baselines. The inclusion of corpus-specific layers in the acoustic model architecture plays a crucial role. The shared component learns key characteristics of Portuguese speech, using both sources of information, while the corpus-specific part focuses on applying these characteristics to the specific characteristics of adults and children, respectively. We also observed that the score gap between the children-alone model and the children's score in the multi-task setting is higher in the English setup. This discrepancy can be attributed to the significantly unbalanced amount of training data in the adult corpus compared to the children's amount of training data. While the gap in the Portuguese setup is only around 10 hours, it reaches approximately 250 hours in the English setup. This observation highlights a limitation of the MTL method, wherein the unbalanced representation can be influenced by the mismatch in corpus sizes. It emphasises the need for careful selection of the different datasets used in the multi-task methodology to mitigate such imbalances.

In the last line, we evaluated training the model on children data using a pre-trained adult model, corresponding the the output model of the first line, as initialisation. This TL enhanced the children's results to 21.53% and 25.36% WER for the English and Portuguese respectively. This configuration emerged as the most effective in our experiments for both English and Portuguese children's ASR. When compared to random initialisation, it is shown that the weights learned in the adult pre-trained model are a beneficial starting configuration and allow the TL to use the learnt patterns to tackle children's speech. Indeed, this TL approach avoids the need for the model to learn these patterns from scratch, which is particularly difficult when using data from a highly variable source like children's speech. As a result, transfer learning may be considered a viable strategy for improving the ASR performance for children's speech. This finding is consistent with the literature on hybrid models [6, 151].

In this study, we conducted a knowledge transfer techniques analysis to improve the results of ASR systems for children in both English and European Portuguese. We corroborate the acoustic mismatch between adult and child speech and the importance of the model to encounter child data and its variability. Our investigations revealed that the TL approach is a promising way to improve low-resource children's speech recognition scores. Furthermore, multi-task learning was found to be helpful in the setting of mixed adult-child ASR acoustic modelling. However, in this study, our primary focus is on the transfer from adults to children. Therefore, the effectiveness of such systems trained using only children's data is not clear. Additionally, while both MTL and TL individually improved the model for children's ASR, we would be interested in exploring the potential benefits of their combined use.

3.4 Combining multi-task and transfer learning using multilingual children data

3.4.1 Motivation

In this section, we present our contribution, where we investigate the potential improvement in the performance of children's ASR for low-resourced languages by leveraging children's resources from various languages. Indeed, in many scenarios, both adult's and children's speech data are limited or even unavailable. To overcome the challenges posed by substantial acoustic variability and data scarcity, we propose a novel approach that uses several small-sized corpora of children from diverse languages. Our study extends conventional multilingual training and TL techniques for hybrid HMM-DNN ASR. We combine these techniques in a meaningful way to use knowledge from heterogeneous data sources. Initially, a multilingual model is trained using a MTL objective, aiming to optimise network parameters for the distinct characteristics of children's speech across multiple languages simultaneously. Subsequently, this trained multilingual model is employed to enhance ASR performance for a target language. Note that the target language may be different from those included in the multilingual training stage. This is achieved through transfer learning, where the knowledge gained from the multilingual model is adapted to the specific characteristics of the target language. In our investigation, we aim to answer the following research question: *Can these approaches be used to efficiently exploit low-resource children's speech data from multiple languages?*

3.4.2 The Multilingual-transfer learning approach

We propose a new approach that combines TL and MTL together for improved acoustic modelling of hybrid HMM-DNN ASR. The proposed approach consists of a two-stage procedure using both MTL and TL that extends the existing techniques since these are usually applied separately. First, a multilingual model trained with a MTL objective aims to optimise specifically the shared part of the network to better model the particular characteristics of children's speech. This is done across multiple languages in parallel. In this work, the model is considered multilingual because all the tasks trained during MTL involve corpora of children from different languages. Secondly, we adapt this multilingual model for a specific children's corpus with TL. The motivation for using TL as a second stage is to take advantage of the robust pre-trained model trained during the MTL phase. Indeed, this pre-trained model has potentially learned cross-linguistic information about children's speech but has also seen more children's data than a model trained in a single language. For this purpose, the acoustic model is divided into two parts: the layers close to the input are shared across all languages, and the top layers, near the output, are language-specific. In other words, there are as many output layers as there are languages, i.e., children corpora. It is worth noting that one can incorporate a new language/task in this second stage by adding a new language-specific output, even if this new language/task has not been seen during MTL training (see Figure 3.2).

Our hypothesis is that the more data from heterogeneous sources seen by the acoustic model in the MTL phase, the better the shared layers could capture the underlying characteristics of children's speech. Then, these learned characteristics can be used effectively, later, by the language-specific layers and during the TL step of the procedure (figure 3.2).

Although the MTL and TL approaches adopted in this work have been used previously in other studies, such as [151] and [158], where they successfully applied MTL using children speaking Mandarin and English, obtaining a relative improvement of 16.96% WER in the English children case, it is clear that the successful performance of this approach in the case of English cannot be expected to generalise to other contexts and languages. Indeed, English is a large-size, resource-rich pluricentric language, which should be seen more as an exceptional case rather than an average representative. It is important to emphasise that there is a need for research that investigates whether these methods, which have already been tested for English, also work in new scenarios, such as low to medium-resource languages with fewer resources than English, like Dutch, Portuguese, Swedish, and German.

Although the approaches adopted in this work have been used previously in other studies, for instance [151] and [158] where they successfully applied MTL using children speaking Mandarin and English, obtaining a relative improvement of 16.96% WER in the English children case, it is clear that successful performance of this approach in the case of English cannot be expected to generalise to other contexts and languages. Indeed, English is a large-size, resource-rich pluricentric language which should be seen more as an exceptional case, rather than an average representative. It is important to emphasise that there is a need for research that investigates whether these methods that have already been tested for English also work in new scenarios such as low to medium-resource languages with fewer resources than English, like Dutch, Portuguese, Swedish and German.

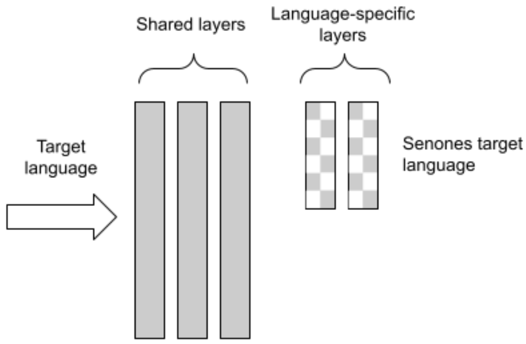


Figure 3.2: Multilingual transfer learning approach. Language-specific layers can be randomly initialised for a language not present during the MTL phase or use the corresponding pre-trained layers in case the target language was present during the MTL phase. Grey blocks are pre-trained during MTL phase.

3.4.3 Experimental Setup

All experiments in this study were conducted using five children corpora, each representing a distinct language: PFSTAR_SWE, ETLTDE, CMU, LETSREAD, and CHOREC. Detailed descriptions of these datasets are provided in Section 2.4. Table 3.3 offers statistics on the duration, number of speakers, number of utterances and language of each corpus.

It is important to note that, in this work, we deliberately focused on utilising small datasets to align with the typical size of available children's speech corpora. Consequently, the MyST children corpus, which is relatively large with over 100 hours of data, was not included in our experiments. Including MyST in this study would introduce a bias toward American English in the multi-task setting, as it represents more than twice the cumulative duration of all our low-resource datasets. While there are ways to mitigate this bias, such as weighting the dataset contributions in the loss function, we chose not to explore this avenue in the scope of this thesis. Our emphasis on smaller datasets aims to better reflect the challenges associated with low-resource scenarios commonly encountered in children's speech research.

For all experiments in this study, we used the same toolkit, architecture, loss functions and learning rate as in Section 3.3.3. We kept the acoustic model divided into two parts: the first part is shared across all languages, and the second part is language-specific. Given that each corpus represents a distinct language, an independent language model and lexicon are employed for each language. Importantly, these language models and lexicons are kept constant across all experiments in this study, ensuring that any observed changes can be attributed to the acoustic model. Additionally, during the backpropagation process, equal weight contributions are given to all datasets.

Corpus name	Language	Train	Test
PFSTAR_SWE	*Swedish*	6030 utterances 138 speakers 4 hours	2879 utterances 60 speakers 2 hours
ETLTDE	*L2 German*	1445 utterances 296 speakers 5 hours	339 utterances 72 speakers 1 hour
CMU	*English*	3637 utterances 76 speakers 6 hours	1543 utterances 75 speakers 3 hours
LETSREAD	*European Portuguese*	3590 utterances 180 speakers 12 hours	1039 utterances 52 speakers 2 hours
CHOREC	*Dutch*	2490 utterances 282 speakers 20 hours	575 utterances 70 speakers 5 hours

Table 3.3: Statistics on the different corpora of children's speech.

3.4.4 Multilingual-transfer learning experiment

Table 3.4 presents the WER results of the MultiLingual Transfer Learning (MLTL) approach compared to three different methods: baseline, trained on each corpus individually for 4 epochs; MTL alone, trained jointly using all corpora for 4 epochs; TL alone, adapted for the target language using, in turn, one of the other 4 baseline models as a source, leading to 4 results per target language. In addition, for clarity, we summarise the transfer learning scores with the average of the 4 scores and the best of the 4 for each target.

Firstly, it is important to emphasise that the baseline scores correctly reflect the different tasks the children were asked to perform and the corresponding amount of data available for each corpus. The best WER score, 21.26% for CMU, can be explained by the reading-aloud-sentences task nature of this corpus. Thus, the language model can more easily compensate for the acoustic model errors. In addition, Chorec and LetsRead, as the largest corpora in our experiment, also yield relatively good results for children's speech recognition. On the other hand, ETLTDE and PFSTAR_SWE show the worse WER results with 44.69% and 54.36% WER, respectively. This can be explained by the limited amount of data available and by the language model which does not compensate as much as the CMU model. Especially for ETLTDE, since it is the only corpus that does not contain scripted text, but spontaneous responses. In addition, the age range of PFSTAR_SWE children also plays a critical role in performance, since younger children generally yield worse performance scores [6].

Regarding MTL, we observed that this approach did not result in improvements in the baseline performance for almost all languages. This observation aligns with the findings discussed in Section 3.3. Nevertheless, we do not observe a significant degradation, suggesting that the model, especially the

	PFSTAR_SWE	ETLTDE	CMU	LETSREAD	CHOREC
Language	*Swedish*	*L2 German*	*English*	*Portuguese*	*Dutch*
Single language	54.36%	44.69%	21.26%	26.88%	25.15%
MTL	54.95%	42.46%	23.01%	27.45%	25.10%
TL from PFSTAR_SWE	-	42.23%	20.62%	26.47%	24.65%
TL from ETLTDE	53.60%	-	20.90%	26.61%	25.42%
TL from CMU	52.83%	41.54%	-	26.49%	24.58%
TL from LETSREAD	52.50%	41.77%	20.41%	-	24.60%
TL from CHOREC	52.20%	40.28%	19.77%	26.05%	-
TL Average	52.78%	41.46%	20.43%	26.41%	24.81%
TL Best	52.20%	40.28%	19.77%	26.05%	24.58%
MLTL	**51.67%**	**38.04%**	**19.33%**	**25.75%**	**23.78%**
MLTL-olo	**51.58%**	40.05%	**19.67%**	26.20%	**24.57%**

Table 3.4: WER results of multilingual-transfer learning and cross-lingual experiments. MTL: Multi-Task Learning, TL: Transfer Learning, MLTL: Multilingual Transfer Learning, MLTL-olo: Multilingual Transfer Learning one-language-out

shared part of the model, is learning shared representations.

Concerning TL, all performance scores surpass their corresponding baseline, confirming that TL is an appropriate method for children's ASR. It allows the system to be exposed to an increased amount of children's data in a non-competitive setting like MTL. Precisely, Table 3.4 indicates that Chorec is the best pre-trained model for knowledge transfer. This aligns with expectations as Chorec is the largest corpus, constituting approximately 40% of the total data used in our experiments.

Finally, MLTL shows an average relative improvement in WER of 7.73% compared to the baseline, slightly higher than the average (TL Avg) and the best (TL Best) TL performance, with an average relative improvement of 4.50% and 2.66%, respectively.

The strength of MLTL is that it can benefit both from MTL and TL, minimising some of their associated weaknesses. Attending to our results, MTL does not improve single language training. We believe that the unbalanced amount of data, the significant differences among data sets and the use of segmental optimisation (LF-MMI) can partially explain these results. This interpretation is further supported by the findings discussed in Section 3.3. Nevertheless, we hypothesise that the multi-task objective leans the network towards better optimisation of the lower layers, rather than optimising the upper language-specific layers, can still be beneficial for TL. Regarding TL, one can observe considerable performance variations depending on the pre-trained model used as the source model, probably due to a poorer initialisation of lower layers that is less efficient for TL. The MLTL experiments show that we can overcome these drawbacks by combining both MTL and TL, thus, validating the effectiveness of this approach for robust speech recognition of children.

3.4.5 Cross-lingual validation

In the previous section, we saw that the MLTL approach yields better results than separate MTL and TL frameworks.

To further validate the hypothesis that the shared lower layers are able to learn meaningful information about children's speech characteristics, regardless of the language, we perform a cross-language experiment following a leave one-language-out cross-validation setting. In this experiment, we keep one language out of the multi-task training and use it only during the TL phase to adapt the acoustic model parameters.

We repeated this procedure for each corpus in our experiment. As in the previous experiment, we used 4 epochs for each learning phase. The last row of Table 3.4 presents the results of the cross-language experiment.

For all corpora, the MultiLingual Transfer Learning one-language-out (MLTL-olo) approach outperforms the baseline WER score with an average relative improvement of 5.56%. Improvements are more important for the small corpora ETLTDE and CMU, with a relative improvement of 14.88% and 9.07%, respectively. PFSTAR_SWE does not benefit as much, with only 5.05% relative improvement. This is mainly due to the age differences with the children in the other corpora used in the MTL phase. Indeed, the children in PFSTAR_SWE are much younger (see section 2.4 for more details). Therefore, we conclude that the shared layers have learned the underlying multilingual features of children.

It is also interesting to compare MLTL-olo with the results of transfer learning alone. In both cases, the pre-trained models used have never seen the target language data. We observe that the results between the MLTL-olo and TL Best are extremely close, with small improvement with the MLTL-olo, only the best transfer learning model on LetsRead is slightly better than MLTL. This means that during multilingual training the system learned, at least, the best representation of the available children's characteristics. This is consistent with our hypothesis of the important role of the multilingual training phase in our two-step procedure.

3.5 Summary and discussion

In this chapter, we have explored the current state of the art for Hybrid HMM-DNN speech recognition systems for children's speech. We aimed to address the following research questions: *Which knowledge transfer approach is best for efficiently modelling and improving automatic recognition of children's speech? Can these approaches be used to efficiently exploit low-resource children's speech data from multiple languages?*.

Our results provide a positive response to these questions, first by demonstrating the effectiveness of the knowledge transfer approach for children's ASR, especially transfer learning. Particularly, we validated the effectiveness of transfer learning in both English and European Portuguese datasets. This efficacy arises from the ability of transfer learning to efficiently leverage knowledge encapsulated in a source pre-trained model trained on adult speech when applied to the task of training on children's speech. In addition, multi-task learning alone may not yield the most optimal results but we have demonstrated that the shared components of the model have the capacity to learn relevant information across multiple tasks jointly. Offering a trade-off for a multilingual or adult-children ASR system. Building upon these insights, this chapter introduces a novel approach—integrating transfer learning and multi-task learning within our multilingual transfer learning system. We demonstrated that the challenges associated with both multi-task and transfer learning can be effectively overcome through the use of our proposed approach. Indeed, this innovative system is built on the strengths of the multi-task learning approach to learn pertinent information jointly, coupled with the efficiency of transfer learning in making effective use of pre-existing knowledge. Even in a low-resource scenario, this approach yielded promising results, resulting in an average relative improvement of 7.73%. Additionally, the benefits of multilingual pre-training extended to transfer learning with an unseen language, showcasing an average relative improvement of 5.56%. These findings underscore the suitability of multilingual transfer learning as a robust method for addressing children's speech recognition challenges, particularly in contexts with limited resources.

In the course of this chapter, our primary focus was directed towards the exploration of a multilingual system; however, it is crucial to emphasise the versatility of our approach. Indeed, our methodology could seamlessly be extended to various other tasks beyond multilingual scenarios. This extends to tasks involving age groups, accents, fluency levels, or varying degrees of intelligibility. It is worth noting that we did not extend our investigation to further scenarios within this thesis, primarily due to the limitations of the available datasets. Nonetheless, the extensibility of our approach holds promise for addressing a wide range of challenges across different linguistic and demographic dimensions in future research. In addition, as a future research direction, it would be interesting to investigate the effects of incorporating a larger children's corpus, adult speech corpus or even non-European languages during the multilingual learning phase with specific weighting on the loss. Additionally, understanding how different types of tasks or linguistic variations influence our approach adaptability would be an interesting avenue of research.

In the context of this thesis, we solely employed a TDNN-F based architecture within the different HMM-DNN systems developed in this chapter. This choice was inspired by earlier research [7] and is acknowledged as the state-of-the-art for children's ASR in HMM-based systems. [7]. However, in response to the emerging use of end-to-end ASR and the promising initial results observed for children's ASR, a strategic decision was made to transition towards the utilisation of end-to-end systems for the remainder of the thesis. This transition underscores an ongoing commitment to staying abreast of advancements in ASR methodologies and exploring innovative approaches that hold the potential to further enhance the recognition accuracy of children's speech.

Finally, it is noteworthy to observe similarities between our approach and the successful XLS-R model [214], which is a self-supervised cross-lingual speech representation learning based on wav2vec 2.0 [163]. Similar to our multilingual transfer learning, both approaches use a two-step procedure where a multilingual model is trained first, and then the model is fine-tuned to a specific language/task. While there are differences, such as dataset size and the supervised/unsupervised settings in the first step of training, recognising potential connections and similarities is valuable. It underscores the relevance and potential impact of our approach in the broader landscape of speech representation learning.

4

End-to-End children automatic speech recognition

Contents

4.1 Introduction

The efficacy demonstrated in incorporating DNN within HMM-DNN setups has motivated a widespread adoption of deep learning techniques in the ASR field. Notably, this increased interest is underscored by the recent achievements obtained with end-to-end models, which have exhibited superior performance compared to traditional hybrid HMM-DNN systems across a diverse array of speech recognition tasks [215, 216]. The primary advantage of end-to-end speech recognition systems lies in the merging of the entire training pipeline within a single neural network, mitigating potential behavioural inconsistencies that may arise between the independent training of the different modules. However, the application of the end-to-end paradigm in the realm of children's ASR is a relatively recent development and has encountered limited exploration [117–120]. Indeed, end-to-end models require a larger amount of data to achieve the desired robustness and flexibility, a condition that is often not met in the case of children's speech. Additionally, the merging of different components within end-to-end models contributes to an increased demand for trainable parameters. Consequently, training end-to-end models on small datasets becomes more challenging [217]. Despite these challenges, the exploration of end-to-end models holds promise for pushing the boundaries of the state-of-the-art in children's ASR.

As discussed in Section 2.2.3, the increased interest in end-to-end speech recognition has led to the development of several architectures, including recurrent neural networks [218], neural transducers [219], and the Transformer architecture [8]. Notably, among these architectures, the Transformer stands out for its ability to consistently yield state-of-the-art results in large-vocabulary speech recognition. This effectiveness is not limited to adult speech but also extends to the domain of children's speech, [117].

Within this chapter, we will present the Transformer design, alongside the Conformer—an improved iteration of the Transformer architecture specifically created for speech-related tasks. Building on prior research that demonstrated the efficacy of Transformer-based models in children's ASR [117], particularly when fine-tuned from pre-trained adult models, our goal is to provide a comprehensive understanding of the various components within these architectures and their roles in achieving optimal TL performances. To this end, we introduce the concept of *partial fine-tuning*, a method where only specific components of the architecture are adapted during TL. The aim of this work is to identify the essential elements underlying the success of transformer-based models during fine-tuning. This knowledge is essential for the further refinement and improvement of methods specifically designed for children's ASR. The objective of this chapter is to address the following research questions: *How do end-to-end automatic speech recognition models achieve state-of-the-art results for children's ASR when finetuned from an adult model? Particularly, what are the components that are most important to fine-tune?*

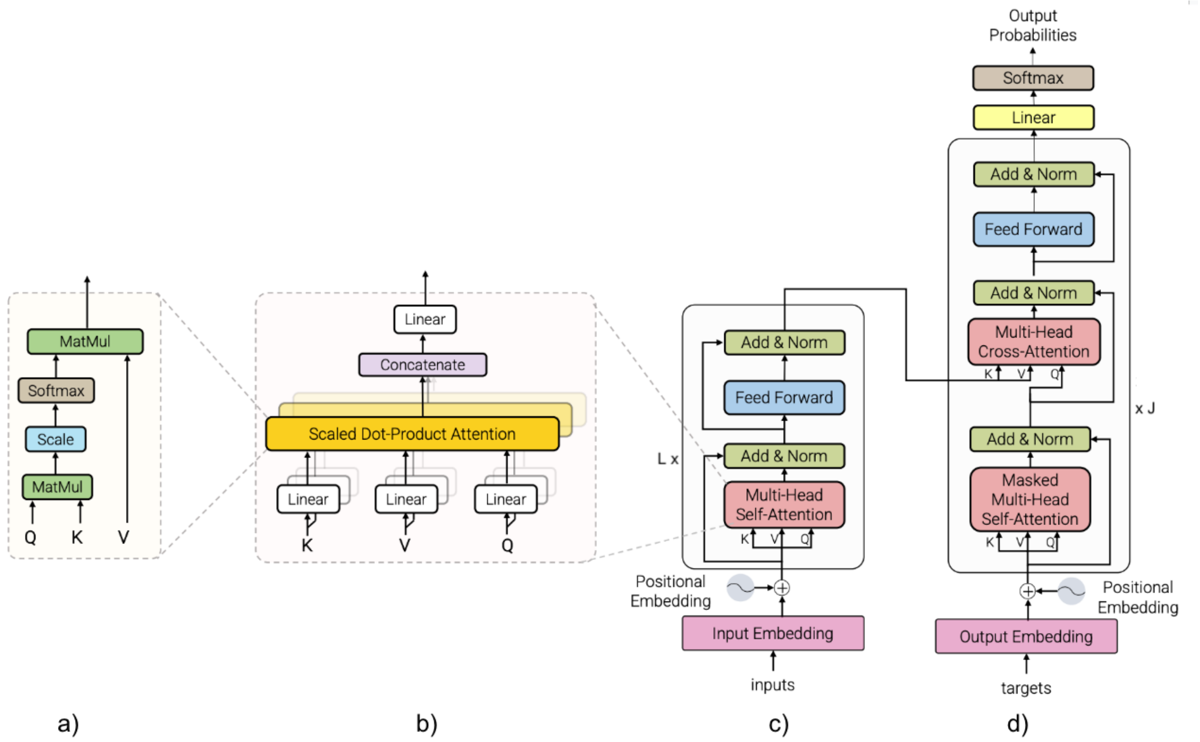


Figure 4.1: Architecture of the standard Transformer [8]. a) scaled dot-product attention, b) multi-head self-attention, c) Transformer-Encoder, d) Transformer-Decoder.

4.2 Transformer model

Introduced in 2017 by Vaswani et al. [8], the Transformer architecture is a sequence-to-sequence Encoder-Decoder model that relies solely on self-attention mechanisms, completely discarding the use of recurrence and convolutions. This design choice addresses challenges such as vanishing gradient issues commonly associated with recurrent neural networks. Another notable difference with recurrent neural networks is that the Transformer computes the dependencies between each pair of positions simultaneously, rather than one by one, by directly encoding the position in the sequence. This enables more parallelisation and therefore a faster training process.

Since its introduction, the Transformer architecture had a tremendous impact across various domains, including NLP [62, 63], computer vision [220], and speech processing [80]. The Transformer's capacity to capture intricate dependencies and patterns in sequences has established it as a popular architecture in the deep learning field, contributing to advancements and breakthroughs across various applications, such as ChatGPT [221] or Dall-E [222].

The Transformer Encoder-Decoder architecture, as depicted in Figure 4.1, consists of an Encoder (c) and a Decoder (d). Prior to entering the Encoder or Decoder, both inputs and targets undergo processing through an embedding layer. This involves the use of learned embeddings to convert input tokens and output tokens into vectors of dimension d_{model}. Since the Transformer model contains no recurrence and no convolution mechanisms, information about the relative or absolute position of the tokens must be injected into the sequence to allow the model to make use of the order of the sequence. To achieve this, information about the relative or absolute position of the tokens is obtained through the summation of the input/output embedding and the positional embedding. While various alternatives for positional encodings were used, Vaswani et al. [8] proposed the use of sinusoidal and cosine functions with different frequencies, as follows:

$$PosEnc_{(pos,2i)} = \sin\Big(pos/10000^{2i/d_{\text{model}}}\Big) \tag{4.1}$$

$$PosEnc_{(pos,2i+1)} = \cos\Big(pos/10000^{2i/d_{\text{model}}}\Big) \tag{4.2}$$

Where pos is the current token or label position and i is the dimension.

The Encoder's primary objective is to transform the input sequence $X = x_1, \ldots, x_T$ into a series of continuous "hidden" representations $Z = z_1, \ldots, z_T$. The Encoder is structured as a stack of L identical layers, each comprising two sub-modules: the Multi-Head Self-Attention (MHSA) and the position-wise fully connected Feed-Forward Network (FFN). Each of these modules is followed by a normalisation with a residual connection.

Subsequently, the continuous "hidden" representations Z are fed into the Decoder. The Decoder is responsible for constructing an output sequence $Y = y_1, \ldots, y_N$ one element at a time. At each time step, the Decoder receives both the Encoder outputs and the last Decoder output in an auto-regressive manner. Similarly to the Encoder, the Decoder is composed of a stack of J identical layers. Nevertheless, in comparison to the Encoder, the Decoder encompasses a third sub-module, which performs Multi-Head Attention (MHA) over the output of the Encoder stack. The self-attention sub-module in the Decoder stack is modified to prevent positions from attending to subsequent positions. This masking combined with a modified MHA prevents the attention from using subsequent positions, ensuring that the prediction at time-step i solely depends on the previous $< i$ time-steps.

The MHA module relies on scaled dot-product attention [8], as illustrated in Figure 4.1(a). Scaled dot-product attention focuses on determining how relevant a particular token is with respect to other tokens in the sequence and is defined as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{4.3}$$

Here, the input consists of queries Q, keys K of dimension d_k, and values V of dimension d_v. The dot product of the query with all keys is divided by $\sqrt{d_k}$, and the result passes through a softmax function to obtain attention weights. The attention weights are then multiplied with the values V. When d_k is large, the scaling $\frac{1}{\sqrt{d_k}}$ restrains the dot product from growing large in magnitude. Note that the MHSA is a specific case of MHA where K, V, and Q are all the same input of the module.

Instead of performing a single scaled dot-product attention, MHA modules linearly project h times K, V, and Q with different, learned, linear projections to dimensions d_k, d_k, and d_v respectively. The attention function 4.3 is then applied in parallel to each of the h projected versions. The output of each of the h attention functions, of dimension d_v, is concatenated and projected one final time, as depicted in Figure 4.1(b). Each of the h attention functions is called a head, while the overall is called MHA or MHSA if K, V and Q are the same. More formally:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{4.4}$$

where

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{4.5}$$

and the different projection matrices are $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$, and $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$.

Furthermore, in addition to the attention modules, each Transformer layer within the Encoder and Decoder encompasses a FFN module. This network is applied to each position separately and identically, and it consists of two linear transformations with a Rectified Linear Unit (ReLU) activation in between. While attention captures interdependencies between the elements of the sequence regardless of their position, the FFN non-linearly transforms each input token independently:

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \tag{4.6}$$

With $W_1 \in \mathbb{R}^{d_{model} \times d_{FFN}}$, $b_1 \in \mathbb{R}^{d_{FFN}}$, $W_2 \in \mathbb{R}^{d_{FFN} \times d_{model}}$ and $b_2 \in \mathbb{R}^{d_{model}}$. Typically d_{FFN} is usually set to $4 \times d_{model}$.

4.3 Conformer model

Transformers are recognised for their effectiveness in capturing global information within sequential tasks, a capability attributed to the attention mechanism. Conversely, CNNs networks excel in capturing local features within data. To leverage the complementary strengths of both architectures, various approaches have been explored [223, 224], and the Conformer architecture [225] stands out as a notable combination

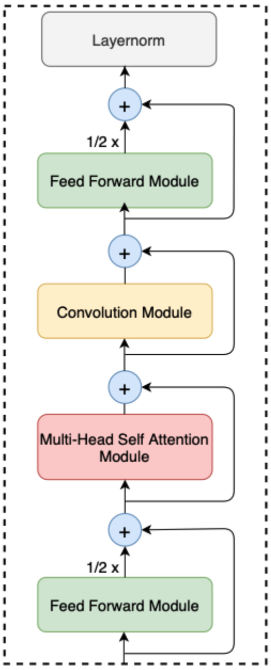


Figure 4.2: Architecture of a Conformer layer

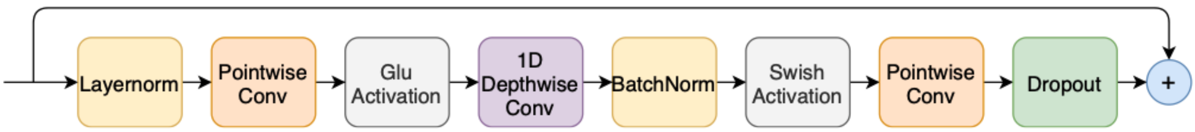


Figure 4.3: Convolution module in the context of a Conformer layer

of Transformers and CNNs.

This combination involves incorporating CNNs into the conventional Transformer architecture, as depicted in Figure 4.2. Specifically, a Conformer block comprises four modules arranged sequentially: a FFN module, a MHSA module, a convolution module, and a second FFN module. Notably, the Conformer block features two FFN modules sandwiching the MHSA module and the Convolution module. This design is inspired by the Macaron-Net [226], which advocates replacing the original FFN in the Transformer block with two half-step FFN modules—one before the attention layer and one after. Similar to Macaron-Net, half-step residual weights are employed for the FFN modules. More formally, for an input x_i to a Conformer block i, the output y_i of the block is defined as follows:

$$
\begin{aligned}
\tilde{x}_i &= x_i + \frac{1}{2}FFN(x) \\
x'_i &= \tilde{x}_i + MHSA(\tilde{x}_i) \\
x''_i &= x'_i + Conv(x'_i) \\
y_i &= LayerNorm(x''_i + \frac{1}{2}FFN(x''_i))
\end{aligned} \tag{4.7}
$$

More specifically, the convolution module used in a Conformer layer is inspired by [227] and illustrated in Figure 4.3. It starts with a gating mechanism [228] involving a pointwise convolution and a Gated Linear Unit (GLU) activation function. Subsequently, a single 1-D depthwise convolution layer is employed. Finally, this 1-D depthwise convolution is followed by a Batch-normalisation and then a Swish activation layer [229].

In summary, Conformers differentiate themselves from Transformers by incorporating convolutional layers to capture local dependencies in addition to the self-attention mechanisms. This design makes Conformers especially suitable for tasks dealing with sequential data, such as speech processing, where the temporal aspect of information is crucial. Notably, this architecture demonstrates improved accuracy with fewer parameters compared to previous approaches on datasets like LibriSpeech [225].

4.4 Understand transfer learning efficacy for Transformer-based models

The end-to-end paradigm has emerged as the state-of-the-art approach for many adult speech datasets, by integrating all components of the traditional HMM-based ASR pipeline, such as acoustic, pronunciation, and language models, into a single neural network. However, this unified neural network design results in a substantial increase in the number of parameters. While handling a large number of parameters is feasible when training on extensive adult data, it poses challenges for smaller children's speech datasets. This limitation has been underscored in [117, 118], where end-to-end models trained on children's speech from scratch were found to be less accurate than traditional hybrid HMM-DNN systems.

In light of this challenge, TL emerges as a promising strategy to address the problem of training large amounts of parameters for a limited-size children's dataset. The key idea is to leverage pre-trained adult models, which are trained on extensive datasets of adult speech. These pre-trained models are adapted to the specific characteristics of children's speech through a re-training phase using the children's dataset. Notably, this adaptation process relies on the use of the knowledge gained during the pre-training phase, eliminating the need to train models from scratch and making the process more manageable for smaller target datasets. The efficacy of the TL approach for children ASR has been demonstrated not only in traditional HMM-DNN approaches, as discussed in the previous Chapter 3 and supported by existing literature [105], but also in modern end-to-end paradigms [117, 118].

With the recent trend of increasing model complexity and parameter count, it becomes crucial to comprehend how these models perform when fine-tuned with limited downstream data. Training models with an expanding number of parameters on a relatively small dataset may potentially result in decreased performance. Consequently, a more detailed exploration of TL for children's speech is needed to identify which parts of the network are more important in the fine-tuning process.

Moreover, the well-recognised issue of overparameterisation in large Transformer-based models, initially discovered in the NLP field, adds an additional layer of complexity. Indeed, models like BERT [62] have been acknowledged to be overparameterised in various studies [230, 231]. Overparameterisation occurs when models have more parameters than necessary for a given task. Empirical observations suggest that certain components or layers of the architecture can be removed without compromising performance, and in some cases, may even lead to slight performance gains [230–232]. In addition, the recognition of overparameterisation has paved the way for successful compression studies, including pruning and distillation techniques [233, 234]. These approaches aim to address the challenges posed by an excessive number of parameters by reducing the model sizes without compromising performances.

In view of the overparameterisation problem, there is a growing need for ablation studies, involving the systematic removal of components of the model, to understand which parts significantly contribute to performance [235, 236]. These studies, predominantly explored in the field of computer vision [232], align with the Lottery Ticket hypothesis formulated by Frankle and Carbin [237]:

> *"A randomly-initialised, dense neural network contains a subnetwork that is initialized such that—when trained in isolation—it can match the test accuracy of the original network after training for at most the same number of iterations".*

While ablation works have been studied in other fields, they remain under-explored for speech tasks. As a matter of fact, the recent successes of distillation and pruning techniques for speech models [238–240] suggest that overparameterisation is also present in ASR models. Consequently, understanding the contribution of the different components of large-size models would not only help optimise model architectures for specific tasks but also reduce the computational demands of training and inference. This is particularly relevant in scenarios with resource constraints, such as limited computational power, memory, and training data. For example, in the context of children's ASR, where data scarcity is a significant challenge, understanding how overparameterisation affects the model could pave the way for the development of more tailored and efficient models.

4.4.1 Partial Transfer learning

In this Chapter, our objective is to conduct a comprehensive exploration of TL, specifically on end-to-end children's ASR. Notably, previous research in this field has solely been focused on HMM-DNN models, as illustrated by the work of Shivakumar et al. [105]. This previous work provided recommendations for the development of better children's ASR systems, but such a study is currently missing for the end-to-end paradigm, which further motivated our investigation. It is noteworthy that existing works on TL on the end-to-end paradigm have solely focused on the entire model fine-tuning, leaving a notable gap in the understanding of the impact of fine-tuning individual components.

Firstly, we perform a meticulous examination of the TL process, specifically isolating the effects on individual components of the Encoder and Decoder, in comparison to the fine-tuning of the entire model in Transformer-based models. The prevailing hypothesis asserts that the Encoder is capturing acoustic information, while the Decoder encodes more linguistic information. Considering the important presence of acoustic variability in children's speech, our investigation extended to discern which layers of the Encoder are more relevant for achieving effective TL and how many of them need to be fine-tuned.

Subsequently, our focus shifts to delineating the distinctive contributions of modules within both the Transformer and Conformer architectures during the fine-tuning process of adapting a pre-trained adult model to children's speech. Our approach called *Partial Fine-Tuning*, aims to investigate the existence of a "Lottery winning ticket" module specifically within the fine-tuning of a Transformer-based model. In contrast to previous work, where the entire model or entire layers are fine-tuned, our granular analysis takes a unique approach by assessing the individual roles of key components, independently of the layers they are in. Namely MHSA, FFN, and normalisation layers for the Transformer and Conformer model in addition to the convolution modules for the Conformer. The objective is to gain insights into the necessity and impact of each module, allowing us to selectively and partially fine-tune the model using a small children's dataset. Therefore, we aim with this approach to require fewer trainable parameters while maintaining the model's effectiveness. Indeed, by departing from the conventional whole-model or whole-layer fine-tuning, our approach aims to optimise the utilisation of limited data resources, contributing to the development of more efficient and tailored models for children's ASR.

4.4.2 Experimental setup

4.4.3 Corpus

Given that the end-to-end paradigm involves encapsulating both the acoustic model and language model within the same network, careful consideration is essential when selecting the dataset for experimentation. Specifically, if the model is trained on a restricted set of repeated prompts, such as a dataset focused on reading tasks, it may learn and overfit to those specific prompts, potentially compromising its ability to

Training	Validation	Test
60897 utterances	10044 utterances	4079 utterances
566 speakers	79 speakers	91 speakers
113 hours	18 hours	13 hours

Table 4.1: My Science Tutor Children Speech Corpus statistics

recognise spontaneous speech and yielding unreliable results. In response to this concern, we decided to use the Boulder Learning My Science Tutor (MyST) corpus, a spontaneous children's speech dataset, as described in section 2.4.

We want to emphasise that we employed the same data filtering and splits as the experiment previously reported using HMM-DNN settings, as detailed in Section 3.3. A summary of the corpus details is recapitulated in Table 4.1.

4.4.4 Implementation details

All experiments were conducted using the SpeechBrain toolkit [241]. The Transformer model encompasses 12 Transformer layers in the Encoder and 6 Transformer layers in the Decoder, all with a hidden dimension of 512. Similarly, the Conformer architecture featured 12 Conformer layers in the Encoder and 6 Transformer layers in the Decoder, with a hidden dimension of 512. Both configurations used 8 heads for all MHSA, a FFN hidden dimension of 2048, and a dropout rate of 0.1. These models were pre-trained on a large English adult speech corpus, specifically the entire LibriSpeech dataset [31], comprising 1,000 hours of data. For reproducibility, these pre-trained models are publicly available[1]. Furthermore, for all experiments, the same Transformer language model was employed, trained on 10 million words from LibriSpeech transcriptions. Our training involved 30 epochs with a learning rate of 8×10^{-5}. Furthermore, in line with findings of [117], a combination of CTC and Seq2Seq losses was used, with respective weights of 0.3 and 0.7.

4.4.5 Encoder-Decoder Transfer learning

Table 4.2 summarises the results of the impact of isolating fine-tuning of the Encoder and Decoder components within Transformer and Conformer ASR models. For the Transformer model, fine-tuning the entire model exhibits a WER of 12.99% using 71.5 million parameters. Isolating the Encoder component leads to a significant improvement in WER, with 12.55% WER with a reduced parameter count of 37.8 million. In parallel, fine-tuning only the Decoder underperformed compared to the full model and Encoder only fine-tuning strategies by achieving a WER score of 15.95% with 25.2 million parameters.

Turning to the Conformer model, the full model achieves a WER of 12.28% with 109 million parameters

[1]https://huggingface.co/speechbrain/ASR-Transformer-Transformerlm-librispeech
https://huggingface.co/speechbrain/ASR-Conformer-Transformerlm-librispeech

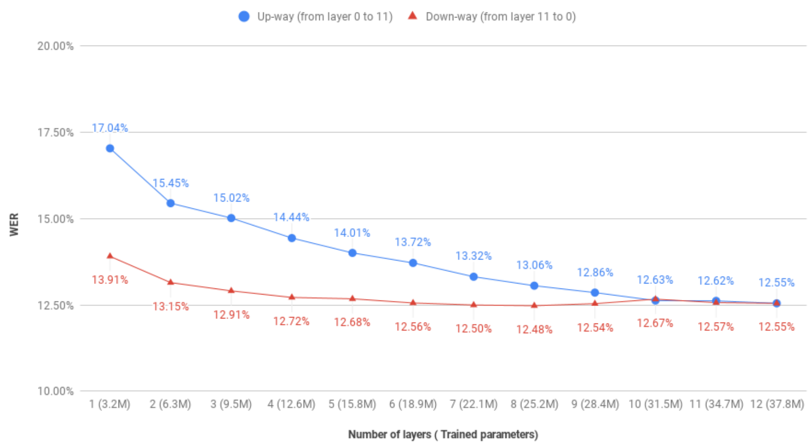
Transformer	WER ↓	Trained parameters
Full model	12.99%	71.5M
Encoder only	**12.55%**	37.8M
Decoder only	15.95%	25.2M
Conformer		
Full model	12.28%	109M
Encoder only	**11.24%**	75.9M
Decoder only	16.94%	25.2M

Table 4.2: Encoder-Decoder experiment

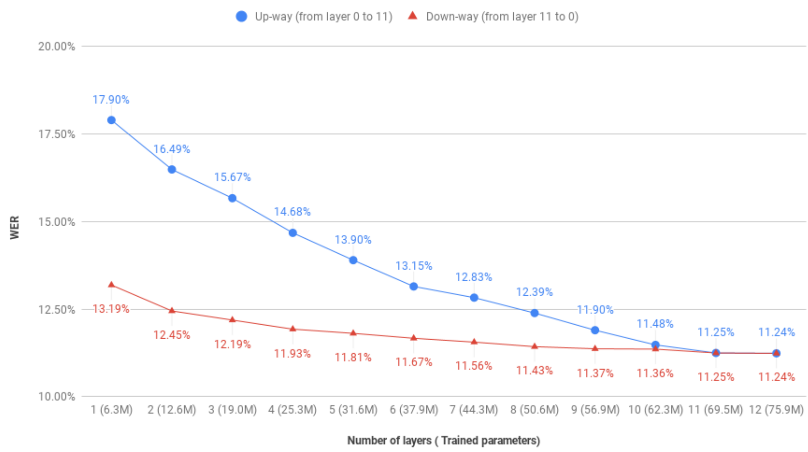
updated. Isolating the TL on the Encoder only yields a remarkable improvement, resulting in a WER of 11.24% with a parameter count of 75.9 million. Conversely, when the Decoder only is fine-tuned the performances degraded with a higher WER of 16.94% using 25.2 million parameters. Notably, the Conformer architecture consistently outperformed the Transformer across all configurations, emphasising its effectiveness for speech-related tasks. In addition, these results underscore the important role of the Encoder in both Transformer and Conformer ASR models compared to the Decoder. Our results highlight the Encoder's role in capturing the inherent variabilities in children's acoustics. We hypothesise that the acoustics variabilities represent the most important source of variabilities as suggested by [6].

Recognising the pivotal role of the Encoder in the success of the fine-tuning process for children's ASR, our investigation delves deeper to determine the specific layers that are the most important during this TL process. To this end, we adopted a meticulous approach where we incrementally fine-tuned the Encoder by adding one layer at a time for each fine-tuning experiment. This layer-wise fine-tuning procedure is executed bidirectionally, encompassing both an up-ward trajectory, commencing from the input layer, and a down-ward trajectory, starting from the output layer of the Encoder. In other words, the up-ward trajectory involves progressively fine-tuning by adding one layer at a time for each experiment, starting from the input layer and systematically integrating subsequent layers towards the output layer of the Encoder. Conversely, the down-ward trajectory initiates fine-tuning from the output layer, systematically incorporating preceding layers towards the input layer at each experiment. The results of this layer-wise fine-tuning procedure for both the Transformer and Conformer architectures are presented in Figure 4.4.

In both the Transformer and Conformer scenarios, a consistent pattern emerged. The addition of more layers was found to be consistently beneficial, with optimal performance stabilisation occurring when 10 layers out of the 12 are employed. We observed that it is more advantageous to use TL from the top layers in the down-ward direction (i.e. those close to the output of the Encoder) compared to the bottom layers in the up-ward direction (i.e. those close to the input of the Encoder). The success of the up-ward trajectory, in contrast to the down-ward trajectory, indicates that the top layers of the Encoder are the most relevant to fine-tuning in the context of children's ASR. This exploration of layer-wise fine-tuning sheds light on the dynamics of TL within the Encoder, offering valuable insights into the targeted layers



(a) Results of the transfer learning layer-wise for the Transformer model



(b) Results of the transfer learning layer-wise for the Conformer model

Figure 4.4: Layers-wise up-way and down-way transfer learning experiment for Transformer and Conformer architecture

Transformer	WER ↓	Trained parameters
Frozen-pre trained	25.04%	-
Full model	12.99%	71.5M
Normalisation	17.00%	57.9K
MHSA	12.19%	25.2M
FFN	**11.84%**	37.8M
MHSA + FFN	12.39%	63.0M
Normalisation + FFN	12.19%	37.9M
Normalisation + MHSA	12.29%	25.3M
Conformer		
Frozen-pre trained	21.75%	-
Full model	12.28%	109M
Normalisation	15.61%	63.7K
MHSA	11.74%	28.4M
Convolution Module	11.67%	9.7M
FFN	**11.10%**	63M
↪ Module 1	11.44%	25.2M
↪ Module 2	11.48%	25.2M
↪ Up-linear (W_1)	11.47%	31.5M
↪ Down-linear (W_2)	11.40%	31.5M
FFN + MHSA	11.20%	91.4M
FFN + Convolution Module	11.11%	72.7M
FFN + Normalisation	11.15%	63.1M
MHSA + Normalisation	11.67%	28.4M
MHSA + Convolution Module	11.44%	38.0M
Convolution Module + Normalisation	11.62%	9.7M

Table 4.3: Modules fine-tuning experiment

that significantly contribute to optimising the performance of children's ASR models.

4.4.6 Modules Transfer learning

The results of our *partial transfer learning* experiments, focusing on fine-tuning specific components of Transformer and Conformer ASR models for children's speech, are presented in Table 4.3. In addition to the WER evaluation metric, we also display the number of trained parameters to evaluate the parameter cost of the different components.

The baseline performance of the Transformer pre-trained model without any fine-tuning (corresponding to the *Frozen-pretrained* line) yields a WER of 25.04%. In contrast, the fine-tuning of the entire Transformer model exhibits a noteworthy improvement, achieving a WER of 12.99% with 71.5 million parameters trained.

The fine-tuning of specific components reveals valuable observations. Applying TL on the normalisation layers alone results in a modest improvement WER of 17.00% by using 57.9 thousand parameters. The fine-tuning of MHSA modules outperforms normalisation and full fine-tuning, achieving a WER of 12.19% with a parameter count of 25.2 million. The most important improvement was observed with the

FFN modules, which attained a remarkable WER of 11.84% using 37.8 million parameters. Remarkably, both MHSA and the FFN modules, when fine-tuned individually, already outperform the full model performance. This implies that the decrease in the number of parameters, coupled with the significance of these modules, may play a substantial role in the enhanced performances of the fine-tuning process with a limited dataset.

Turning to the Conformer model, the baseline WER without fine-tuning gives 21.75%. In contrast, the full fine-tuning of the Conformer model yielded an improved WER of 12.28% with 109 million parameters.

Fine-tuning specific Conformer modules offers further granularity. First, the normalisation layers fine-tuning, in a similar way as observed in the Transformer configuration, yield a score of 15.61% WER, with 63.7 thousand parameters. Then, MHSA modules proved to be effective by already providing better results than the full full-tuning with a WER of 11.74%, by training 28.4 million parameters. The convolution modules outperformed the MHSA with a WER of 11.67% with fewer parameters used, 9.7 million. However, as in the Transformer model, FFN modules stand out significantly, demonstrating a WER of 11.10% with a parameter count of 63 million. Notably, all the MHSA, convolution modules and FFN modules, when fine-tuned in isolation, surpass the performance of the full Conformer model.

As the FFN modules consistently proved to be the most relevant component to fine-tune for children's ASR, irrespective of the configuration, we opted to delve deeper into the FFN submodule. We identify two ways to subdivide the FFN components of a Conformer model. The first split is along the macaron-style of a Conformer FFN layer, which includes two modules—one before the MHSA and one after the convolution module, referred to as Module 1 and Module 2, respectively. The second subdivision of a FFN module involves focusing solely on the up-linear and down-linear components, denoted as W_1 and W_2 in Equation 4.6. Using the first splitting, Module 1 and Module 2 achieve WERs of 11.44% and 11.48%, respectively, each by fine-tuning 25.2 million parameters. While the Up-linear and Down-linear submodules exhibit WERs of 11.47% and 11.40%, respectively, with parameter counts of 31.5 million. The subdivision of the FFN modules did not result in better performance than their coupled usage. This underscores the significance of fine-tuning the full FFN modules in Transformer-based end-to-end models.

Furthermore, our investigation delved into the impact of combining different components within both the Transformer and Conformer architectures. In the case of the Transformer model, the results consistently showed that while all combinations improved upon the 12.99% WER achieved through full model fine-tuning, they consistently fell short of the score achieved by fine-tuning the FFN components alone, which was 11.84% WER. The most promising combination attained a WER score of 12.19%, by using normalisation layers and FFN modules in combination. Similarly, within the Conformer architecture, a comparable trend emerged. While all combinations exhibited improvements compared to the full model fine-tuning score of 12.28%, they still lagged behind the performance achieved with the FFN-only scenario of 11.10%. A noteworthy result was the observation that the combination of the FFN and convolution modules proved to be as effective as the FFN in isolation, yielding a WER score of 11.11%. This particular experiment accentuates the notion that employing a single component is more advantageous than utilising combinations, thereby suggesting the benefits of a more parsimonious use of parameters while fine-tuning on a small dataset.

This observation is consistent with the "Lottery Ticket Hypothesis," which suggests that within a neural network, there exist subnetworks (winning tickets) that, when isolated, can perform equally well as the full network. In the context of TL for children's end-to-end ASR, our findings emphasise the role of the FFN modules as winning tickets.

4.5 Summary and discussion

In this chapter, we delve into the end-to-end paradigm for children's ASR, particularly using transfer learning from a pre-trained adult model. We aimed to answer the following research questions: *How do end-to-end automatic speech recognition models achieve state-of-the-art results for children's ASR when finetuned from an adult model? Particularly, what are the components that are most important to fine-tune?*

In contrast to prior research where fine-tuning involved the entire network, our detailed evaluation of different levels of fine-tuning provides insights into the underlying behaviour of the model when adapted for children's ASR. This research offers valuable recommendations for the future development of children's end-to-end ASR systems. Initially, we identified the Encoder as the most crucial part of the model to fine-tune compared to the Decoder. This aligns with the hypothesis that substantial variabilities in children's speech are present in the acoustics and can be effectively captured in the Encoder, echoing similar findings in transfer learning experiments for HMM-DNN framework in [6].

Addressing the second research question, our proposed partial fine-tuning procedure sheds light on the most important components for adaptation in both Transformer and Conformer architectures. This experiment unveiled the presence of the Lottery Ticket Hypothesis in such end-to-end architectures, indicating overparameterisation. Through our partial fine-tuning approach, the FFN module emerged as the most critical component to fine-tune, our winner ticket, outperforming all other configurations, even the entire model fine-tuning. Previous research [242] has revealed that FFN layers exhibit key-value memory properties, offering a plausible explanation for their success in the transfer learning procedure. Moreover, we observed that combining these submodules can potentially deteriorate ASR results, underscoring the challenge of over-parameterisation and the need for training on a smaller amount of parameters when using a small dataset.

As the ASR field advances, one notable trend is the ever-increasing size of models, where parameters in neural networks are reaching unprecedented scales. Within these expansive architectures, the FFN components often constitute a substantial portion of the model's overall parameter count, accounting for 52% and 57% of the total number of parameters in our Transformer and Conformer architectures, respectively.

The widespread adoption of massive models has raised concerns regarding computational efficiency, optimal resource utilisation, and the potential risk of overfitting when these models are fine-tuned with limited data. Given that FFNs modules constitute a substantial portion of these large-scale models, there is an increased interest in developing parameter-efficient transfer learning strategies. Efficient utilisation of parameters during training is not only crucial for mitigating computational costs but also for enhancing the practicality of deploying these models in real-world scenarios where limited resources may exist. The next chapter of this thesis will delve into the feasibility of employing such parameter-efficient approaches for children's ASR. Indeed, such a parameter-efficient approach could be highly beneficial in the development of speaker-based ASR systems for children.

5

Exploring Parameter-Efficient Strategies in Transfer Learning for Children-Focused ASR Systems

Contents

5.1 Introduction

The use of increasingly larger models coupled with the abundance of massive datasets is driving rapid advancements in many domains of machine learning, encompassing NLP [63] and computer vision [222]. In the context of ASR, this trend of scaling up models is exemplified by state-of-the-art models such as Whisper [30] and HuBERT [164], where the number of parameters can exceed 1 billion. Research has underscored the interconnected nature of the training dataset size and the number of model parameters, identifying them as mutual bottlenecks that influence the performances of machine learning models [243]. This observation accentuates the significance of scaling these two dimensions in tandem for the development of more robust and effective ASR models. Typically, to scale up the model size a combination of an increased number of layers and an expansion of the model's hidden dimensions is used [244].

However, the challenge arises when only a limited amount of data is available, making it challenging to train these large models from scratch, as highlighted by recent studies [117,118]. Hence, as discussed in the preceding chapters, TL emerges as a well-established and effective paradigm to tackle to problem of limited dataset. Nevertheless, despite its efficacy, we emphasised certain limitations that may potentially impede the performance of fine-tuning. Specifically, attempting to fine-tune these large models using a downstream dataset limited in size can be challenging, as shown in Chapter 4. Indeed, in addition to being an expensive process, using a small amount of data on such a large model could potentially result in overfitting. This issue necessitates careful consideration, particularly in light of the recent evolution towards ever-growing pre-trained model sizes. Additionally, even following the last chapter's findings where only specific parts of the model were fine-tuned, the different parts of the model are intricately linked to the overall model size. For example, FFN modules usually represent a substantial portion, between 50% to 70% of the total number of parameters. This insight underscores the persistence of the challenge associated with model size, even when fine-tuning only specific components. Finally, TL on large amounts of parameters is memory-storage-inefficient, especially when there is a need to store replicas of all the models's parameters for many different small tasks.

Consequently, there is a growing need for more Parameter-Efficient Transfer Learning (PETL) as lightweight alternatives to TL. Among the approaches introduced by the research community, residual Adapter modules stand out as the most popular and promising [245, 246]. Specifically tailored for Transformer-based systems, Adapters integrate a compact set of additional layers into a pre-trained frozen source model. This design enables Adapters to enhance computational efficiency, resulting in faster training and addressing the challenge of catastrophic forgetting. Diverging from conventional TL methods, where the source pre-trained model's weights are entirely replaced, Adapter-transfer maintains the integrity of the backbone model. Therefore, when the Adapter modules are removed, the initial pre-trained model remains unchanged. This preservation of the backbone model is a crucial advantage as it offers increased flexibility. Furthermore, owing to their limited number of trainable parameters, Adapters

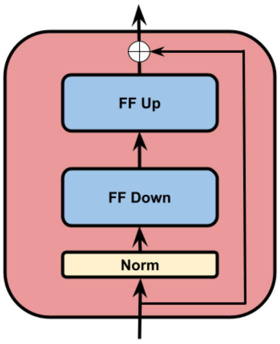


Figure 5.1: Residual Adapter architecture

demonstrate a decreased susceptibility to overfitting, thereby contributing to improved generalisation performance for smaller training datasets.

In this chapter, motivated by the promising characteristics of Adapters, we will investigate their application in the specific context of children's ASR. The study will encompass the examination of diverse Adapter configurations within both Transformer and Conformer architectures. The investigation aims to unveil the potential for creating a model that optimally balances parameter efficiency and recognition accuracy. Additionally, we propose a novel approach, where speaker-group-based Adapters are trained using unsupervised clustering over speaker embeddings. The primary objective of this chapter is to address the following research question: *Is it possible to develop a parameter-efficient automatic speech recognition model for children?*

5.2 Adapter tuning

Adapters were initially introduced in the NLP field to efficiently adapt large models, such as Transformers, using a minimal amount of parameters for text classification [245]. As an alternative to full model fine-tuning, Adapter-transfer involves training an extra small number of task-specific parameters while keeping the original model frozen. To this end Adapters are plugged at each Transformer layer level after the MHA and FFN modules. This setup is often referred to as the *Houlsby* configuration. Subsequently, [246] demonstrated that Adapters placed only after the FFN modules were sufficient for achieving efficient performances, referred to as the *Pfeiffer* configuration. Typically, Adapters use a bottleneck architecture, consisting of a normalisation layer followed by a projection-down linear layer with a non-linear activation, projecting the input into a d_{hidden} dimension. Subsequently, a projection-up linear layer brings back to dimension d_{model}. Finally, a residual connection is applied by summing the input of the Adapter with its output. The overall structure is illustrated in Figure 5.1. Research suggests that the hidden dimension, between the down and up projection, may not always benefit from a bottleneck structure, where $d_{hidden} < d_{model}$, and the optimal design may vary depending on the downstream task [245]. In some tasks, a hidden dimension larger than the model size itself, in other words, $d_{hidden} > d_{model}$, has been proven more effective [171].

Mathematically, the structure of an Adapter can be expressed as follows:

$$adapter(x) = x + (W_{up}(f(W_{down}g(x) + b_{down}))) + b_{up}) \tag{5.1}$$

Where W_{down} and W_{up} denote the weights of the projection-down and projection-up linear layers with respective dimensions of $\mathbb{R}^{d_{model} \times d_{hidden}}$ and $\mathbb{R}^{d_{hidden} \times d_{model}}$, and b_{down} and b_{up} represent the corresponding biases. The function $f(\cdot)$ is a non-linear activation, while $g(\cdot)$ is a layer normalisation or identity function. Finally, x corresponds to the input given to the Adapter.

In terms of computation, Adapters offer the advantage of faster training, given that they update fewer parameters compared to fully fine-tuning models. Nevertheless, there might be a slight processing delay during inference due to the introduction of extra parameters by the Adapters; however, this difference is generally minimal and can be well-managed [247].

Recently, a rising interest has been observed in Adapter-transfer for ASR tasks [248–250], particularly owing to its modular nature, which has proven advantageous in the context of multi-lingual ASR [251–253]. In these studies, distinct Adapters were trained for each language, contributing to enhanced performance compared to a monolingual model and mitigating certain challenges associated with TL, such as overfitting. This modular approach provides a tailored solution, as each Adapter designed for a specific language can effectively capture the diverse acoustic characteristics unique to that language.

Moreover, researchers have explored the use of Adapters in the context of SSL. Typically, in SSL, larger models are employed to capture a diverse range of information from speech, applicable across a broad spectrum of tasks [171,254]. However, the computational cost and scalability to adapt these models for multiple tasks can be challenging. Notably, once the model is fine-tuned for a specific task, the entire model is fixed for that task, and re-loading and re-training the base model are necessary for transferring to a different task. Therefore, the use of Adapters has proven effective in addressing these challenges by providing a modular and parameter-efficient task-specific adaptation.

Finally, the effectiveness of Adapters has also been demonstrated in addressing challenges related to low-resource and atypical speech recognition scenarios [255]. This underscores the adaptability and robustness of Adapters, particularly in scenarios where data may be limited or exhibit unconventional characteristics.

However, the application of Adapters in the context of children's ASR has received limited attention, with only one notable study by [171]. In this study, the authors introduced a novel approach that involved integrating and training Adapters within SSL setting, followed by fine-tuning the entire model, including the Adapter weights, to enhance the modelling of children's speech. This represents a pioneering effort

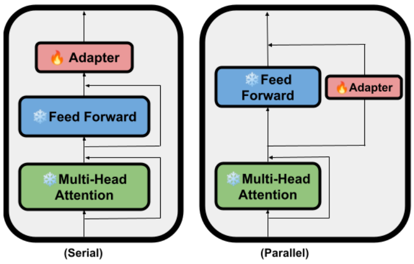


Figure 5.2: Transformer block with various residual adapter configurations (Normalisation layers are not shown in this picture for simplicity in plotting.)

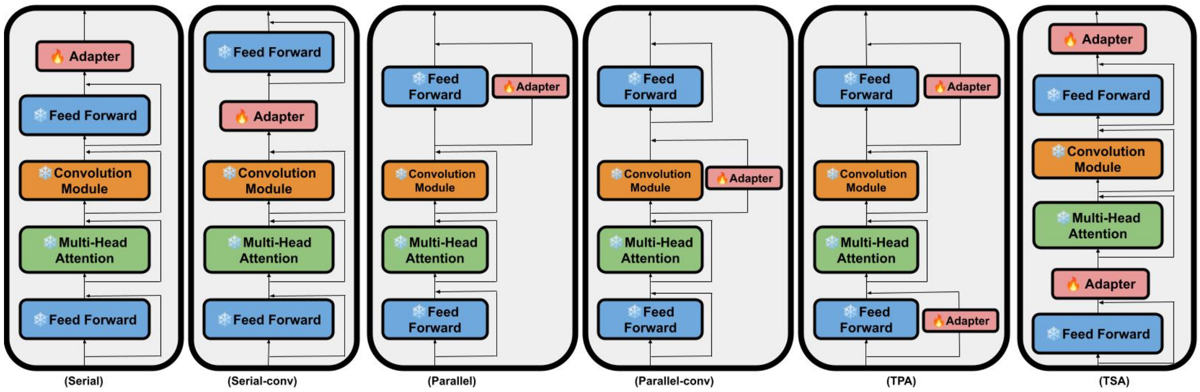


Figure 5.3: Conformer block with various residual Adapter configurations (Normalisation layers are not shown in this picture for simplicity in plotting.)

to leverage Adapters for adapting large-scale models to the unique characteristics of children's speech. Nevertheless, it is crucial to note that in this work, the authors updated the entire model along with the Adapter, compromising the parameter efficiency associated with Adapter-transfer. This highlights the need for a more focused investigation into Adapters as a PETL method for children's ASR.

5.3 Investigating Adapters for Children's ASR

In this section, we delve into the application of Adapter-transfer as PETL for both Transformer and Conformer architectures in the domain of children's ASR. Building upon the insights gained from our partial fine-tuning approach in Chapter 4, we identified FFN modules as the most relevant components for fine-tuning in a Transformer-based model. Consequently, we choose to use Adapters for modifying the output of these FFN modules. Additionally, given our results that underscored the significance of fine-tuning the Encoder, our primary focus will be on exploring the application of Adapters within the Encoder. For the Transformer architecture, we explore two methods of integrating Adapter modules into the model: *parallel* and *serial* placement with each FFN component. These two configurations were used in prior work [256] and are depicted in Figure 5.2.

In the case of the Conformer architecture, we explore six distinct Adapter configurations, as illustrated in Figure 5.3. The initial two configurations mirror our Transformer investigation, involving both *parallel* and *serial* placements, either after or in parallel with the second FFN module [249]. Additionally, we assess a configuration that introduces an Adapter following the convolution module, denoted as the *serial-conv* setup used in [250]. Notably, although the FFN component has been identified as the most crucial for fine-tuning, promising results with our partial fine-tuning have been observed by fine-tuning the convolution modules only. Furthermore, [249] introduces two variants of the parallel setup: *parallel-conv* where the Adapter operates in parallel with the convolution module, and the *Two Parallel Adapters (TPA)* configuration where two Adapters are placed in parallel with both FFN modules in each Conformer layer. To comprehensively explore all feasible configurations, we introduce a novel configuration, the *Two Serial Adapters (TSA)* where two Adapters are sequentially positioned after each FFN component in all layers. This comprehensive exploration of Adapter configurations within both architectures aims to discern the most effective adaptation strategies for children's speech in the context of ASR.

Mathematically, in serial configurations, the Adapter input is provided by the preceding component denoted as P, which can be either FFN or convolution module, depending on the specific configuration. The output of the Adapter is then determined by the following process:

$$output = Adapter(P(x)) \tag{5.2}$$

where x represents the input of component P.

In parallel configurations, the process varies slightly. In this scenario, the Adapter's input is the same as P, and the Adapter's output is combined with the output of component P as follows:

$$output = x + 0.5 \cdot P(x) + (Adapter(x) - x) \tag{5.3}$$

Furthermore, we consider three distinct configurations where Adapters are integrated into the Decoder. It is important to note that in the Conformer architecture, the Decoder consist of regular Transformer layers. Consequently, we assess both the *Serial* and *Parallel* Adapter setups. Additionally, we examine the combination of the most effective Encoder and Decoder Adapter configurations. To the best of our knowledge, there is no prior research that formally investigates the influence of Adapters within a ASR Decoder.

Finally, motivated by the observed strong correlation between children's speech variability and age [6], we explore the possibility of training specialised Adapters. However, considering the individual growth speed of each child may not align with a predefined age, using age groups directly may not effectively capture children with similar acoustic characteristics. Additionally, in many children's speech datasets, age information is often not provided. To address this, we propose to partition the dataset into groups of speakers with similar acoustic characteristics based on unsupervised clustering of speaker embeddings. In practice, we apply a k-means clustering algorithm on the x-vector representation [257] of all training utterances. Subsequently, distinct Adapters are trained for each speaker cluster separately. During the testing phase, the closest cluster of the group of speakers is determined for each test utterance, and the corresponding Adapters specific to that group are employed for decoding. The primary objective of these experiments is to investigate whether Adapters trained on comparable speech characteristics yield improvements over a general Adapter on the entire training set.

5.4 Implementation details

For all experiments, we used the same pre-trained Transformer and Conformer models described in Section 4.4.4. The architecture of each Adapter comprises an initial linear layer projecting to dimension 512 with a ReLU activation, followed by another linear layer projecting to dimension 512 with a residual connection from the Adapter input. In the initialisation process for all Adapters, W_{down} was set to all zeros, and W_{up}, b_{down}, b_{up} were initialised using Xavier initialisation [258].

The decision to use a hidden dimension size (d_{hidden}) equal to d_{model} instead of employing a bottleneck was influenced by prior research on hidden dimension size. Previous studies consistently demonstrated that larger dimensions tend to result in improved performance scores [249]. All models underwent training of 30 epochs, with a learning rate of 8×10^{-4} for training the Adapters and 8×10^{-5} for fine-tuning the entire model. In the clustering experiments, we applied the k-means clustering algorithm to the speaker embeddings of each utterance. The speaker embeddings were extracted using a publicly pre-trained ECAPA-TDNN model, trained on adult speech[1].

For training, we use the My Science Tutor (MyST) Children Speech Corpus, as children dataset. The Myst dataset is the same as it was employed in prior experiments conducted on this dataset within the context of this thesis, as detailed in Table 4.1. For a more comprehensive understanding of the MyST corpora, additional details are provided in Section 2.4.

5.5 Results

5.5.1 Configurations

In this section, we present a comprehensive evaluation of the different Adapter configurations applied to both Transformer and Conformer models. These results are presented in Table 5.1. First, we assess the Transformer model when no fine-tuning was applied (*Frozen*), resulting in a WER of 25.04%. Conversely,

[1] https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

Table 5.1: Results of the different Adapters configurations in both Transformer and Conformer.

Method	WER ↓	Trained params
Transformer		
Frozen	25.04%	-
Full fine-tuning	12.99%	71.5M
Serial	12.78%	6.3M
Parallel	**12.62%**	6.3M
Conformer		
Frozen	21.75%	-
Full fine-tuning	12.28%	109.1M
Serial	11.76%	6.3M
Serial-Conv	11.78%	6.3M
Parallel	11.72%	6.3M
Parallel-conv	11.79%	6.3M
TPA	**11.58%**	12.6M
TSA	11.75%	12.6M
Serial (Decoder)	18.09%	3.2M
Parallel (Decoder)	17.76%	3.2M
TPA + Parallel (Decoder)	**11.47%**	15.8M

Full Fine-Tuning involved complete fine-tuning of the entire model, working as our baseline system, reducing the WER significantly to 12.99%, with the use of 71.5 million trainable parameters. Turning to the Adapter setups, we investigate the *Serial* and *Parallel* configurations, both using 6.3 million trainable parameters. The *Parallel* emerged as the best configuration, achieving the lowest WER of 12.62% compared to 12.78% for the *Serial*. These results underscore the effectiveness of Adapter configurations within the Transformer architecture, as they both perform slightly better than the full-finetuning baseline.

Next, we investigated the Conformer architecture, we once again explored *Frozen* and *Full Fine-Tuning*. The *Frozen* pre-trained model yielded a WER of 21.75%, while the full fine-tuning, in a similar way as the Transformer, led to enhanced performance, with WER of 12.28% using a total of 109.1 million trainable parameters. Within the set of Adapter configurations, *Serial* achieved a WER of 11.76%, while *Parallel* demonstrated slightly better performance with a WER of 11.72%. These results confirm that *Parallel* Adapters were more effective in improving WER in both Transformer and Conformer models. When Adapters are placed after the convolution layer, with the *Serial-conv* and *Parallel-conv* configuration, both slightly under-perform compared to Adapters placed after the second FFN component with respective scores of 11.78% and 11.79%. Finally, we evaluated the *TPA*and *TSA* configurations. The *TPA* configuration emerged as the most promising, with a remarkable WER of 11.58% using 12.6 million trainable parameters, while *TSA* achieved a WER of 11.75%, which is slightly under-performing compared to the *TPA* configuration.

In addition, we evaluated the use of Adapters in the Decoder. As the Decoder of the Conformer

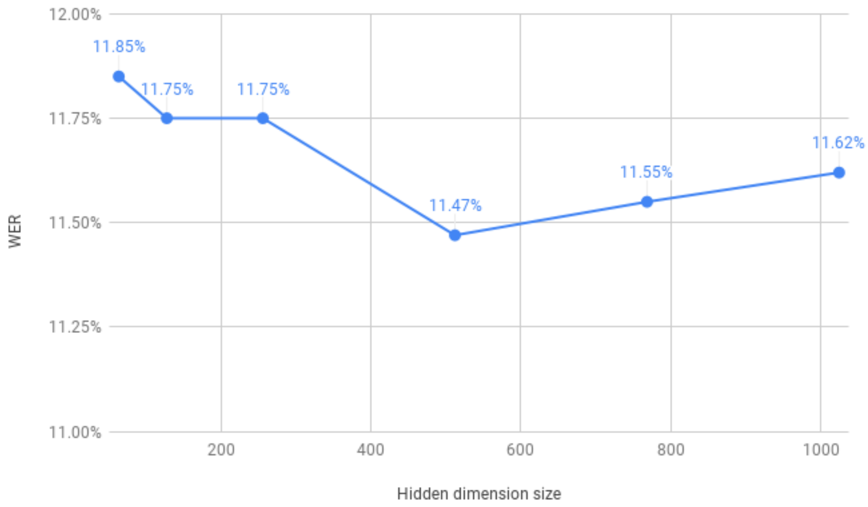


Figure 5.4: Experimental Adapter transfer using different hidden dimension sizes within the Conformer architecture

architecture is a regular Transformer, we only evaluate the *Serial* and *Parallel* setup, which respectively reached 18.09% and 17.76% WER with 3.2 million parameters. These results showed that Adapters are more relevant when plugged into the Encoder which is in line with findings from Chapter 4. Finally, combining *TPA* in the Encoder layers with *Parallel* Adapters in the Decoder outperforms Adapters in the Encoder only, with 11.47% WER. Consequently, this configuration stands as the most effective configuration for our children's speech dataset.

We performed statistical tests (Matched Pairs Sentence-Segment Word Error) across all Adapter setups in comparison to the full fine-tuning configuration using SCTK, the NIST Scoring Toolkit [2]. The results reveal that, in all scenarios, the p-value is less than or equal to 0.001. This observation denotes statistical significance, indicating evidence against the null hypothesis.

These results collectively illustrate the versatility and effectiveness of different Adapter configurations within the Transformer and Conformer model for the children's ASR task. *TPA* Adapters in the Encoder combined with *Parallel* Adapters in the Decoder showcased outstanding performance, highlighting their potential as a fine-tuning replacement in large model children ASR scenarios.

5.5.2 Effect of the Adapters hidden dimension

In this section, our objective is to assess the influence of different hidden dimension values for the Adapter (d_{hidden}) on the model's performance. In our previous experiments, we maintained a fixed hidden dimension, where $d_{model} = d_{hidden} = 512$. Now, we aim to explore various hidden dimension

[2] https://github.com/usnistgov/SCTK/tree/master

# of clusters	Average WER ↓
1	11.58%
2	**11.50%**
3	11.57%
4	11.51%

Table 5.2: Results of the unsupervised clustered Adapters approach.

values within the Conformer architecture, focusing on the *TPA* in the Encoder-only configuration. This investigation aims to assess how variations in the hidden dimension may impact the performance of the Adapter transfer, offering valuable insights into the parameter tuning process for Adapter modules.

In Figure 5.4, the WER performances of the Adapter transfer are presented in relation to the hidden dimension size (d_{hidden}). Notably, a trade-off is observed, emphasising the importance of choosing an optimal hidden dimension size. Configurations with dimensions that are either too small or too large result in a degradation of overall performance. The best-performing configuration aligns with the choices made in all previous experiments, with $d_{hidden} = d_{model} = 512$. We hypothesise that the parallel Adapter functions as an extension of the key-value memory of the FFN. Opting for an extremely large hidden dimension makes training more complex due to a large number of parameters, while an excessively small size drastically limits the potential information learned from the Adapters.

5.5.3 Unsupervised clustering for grouped-speaker Adapters

In this section, we present the outcomes of our clustering approach summarised in Table 5.2. The investigation focuses on the influence of varying the number of clusters on the ASR scores, ranging from 1 to 4 clusters, using the *TPA* configuration in an Encoder-only setup within the Conformer model.

Initially, when the data remained unclustered, corresponding to one cluster, the ASR system exhibited a WER of 11.58%. Notably, the two-cluster configuration outperformed the other setups, achieving superior performance with a WER of 11.50%. This result suggests that partitioning the data into two distinct clusters allows the different Adapters to more effectively capture underlying patterns intricately linked to their respective clusters, consequently enhancing the recognition scores.

Furthermore, we explored the impact of increasing the number of clusters to three and four, revealing only marginal differences in performance. Specifically, the three-cluster configuration yielded a WER of 11.57%, and the four-cluster configuration resulted in a WER of 11.51%. These findings underscore the role of data clustering in children's ASR systems by grouping shared speaker characteristics into different clusters.

In summary, our investigation emphasises the crucial role of data clustering in the context of children's ASR systems. Specifically, for the Myst dataset, optimal performance was attained with a two-cluster configuration, suggesting that this approach facilitates the effective capturing of cluster-specific patterns by the Adapters. The marginal performance differences observed with three and four clusters suggest a potential saturation point, indicating that further partitioning may yield diminishing returns in terms of ASR performance improvement for this specific dataset.

It is noteworthy that, in this experiment, the amount of available training data for Adapters varies due to the clustering process. A more comprehensive exploration of the impact of training hours on Adapters will be presented in Section 7.3.3.C. This forthcoming analysis will offer a detailed understanding of how different training durations influence the performance of Adapters in diverse conditions, shedding light on their adaptability and effectiveness across various scenarios.

Additionally, considering the relatively narrow age range of the Myst corpus, encompassing children from the third to fifth grade, future work could explore the applicability of this approach on children datasets with a more extensive age range. This extension would contribute valuable insights into the generalisation and robustness of the clustering-based Adapter approach across diverse age groups within the realm of children's ASR.

5.6 Summary and discussion

In this chapter, we investigated the viability of Adapter-transfer in the context of children's ASR. Addressing the research question, *Is it possible to develop a parameter-efficient automatic speech recognition model for children?* our investigation yielded an affirmative response. Our study showcased the effectiveness of adapting the model using Adapter modules, resulting in improved WER performances compared to full-model fine-tuning. Notably, this was achieved while utilising only approximately 10% of the parameters required in the traditional transfer learning of the entire model, underscoring the parameter efficiency of Adapter-transfer.

Among the various configurations explored in this chapter, the "parallel" Adapter and its extension, the "TPA)" emerged as the most effective choices for the Transformer and Conformer architectures, respectively. The "parallel" Adapter, in particular, demonstrated its efficacy as it extends the key-value memory that represents the FFN modules [242], showcasing its potential to capture essential information specific to children in a more parameter-efficient manner.

Building on these findings, we proposed the integration of unsupervised clustering on the speaker embeddings extracted from different utterances into the Adapter-transfer procedure. This strategic clustering aimed to facilitate the training of specific Adapters for each cluster, offering a more personalised adaptation without relying on detailed metadata about the speaker, such as their age. Our results illustrated that leveraging these clusters could further enhance overall performances, opening avenues for the application of Adapters for better personalised ASR systems.

The successful implementation of Adapter transfer in the domain of children's speech presents promising opportunities for advancing children's ASR. Our findings indicate that Adapters serve as effective tools for bridging the gap between a source and a target domain while retaining the valuable knowledge encapsulated in pre-trained models. This promising outcome lays the foundation for further exploration. In the upcoming chapter, we will extend our research by investigating the application of Adapters in the domain of TTS data augmentation. This extension aims to leverage the capabilities of Adapters to reduce the disparity between real and synthetic children's speech.

Moreover, in addition to the Adapter module introduced in this chapter, it is noteworthy to mention that various PETL alternatives exist in the literature, demonstrating their effectiveness in tasks beyond children's ASR, such as NLP tasks demonstrating in some scenario better accuracy and parameter efficiency. Therefore, we will extend our exploration of PETL modules, evaluating their applicability and performance in the specific context of children's ASR. Additionally, we will delve into the development of new PETL approaches, aiming to strike a better balance between parameter efficiency and accuracy.

6

Integration of synthetic speech for data augmentation

Contents

6.1 Introduction

The continuous progress of deep learning, fueled by the abundance of extensive training datasets, has significantly improved the performances of ASR. Nevertheless, despite these notable advancements, the recognition of children's speech poses a unique set of challenges, leading to performance disparities when compared to adult speech recognition. The inherent variability in children's speech, influenced by factors such as age, linguistic development, and articulatory differences, necessitates the acquisition of a sufficiently diverse and comprehensive dataset for the effective training of children's ASR systems.

In an attempt to bridge this performance gap, researchers in [39] addressed the challenge by leveraging an in-house large dataset of children's speech, comparable in scale to an adult corpus, to train an ASR model. The results demonstrated state-of-the-art performances, highlighting the potential of ASR systems to effectively learn from diverse and variable children's speech data when provided with a substantial amount of training data. However, it is important to note that, in most cases, the creation of such comprehensive datasets for children's speech faces practical constraints, including ethical considerations, privacy concerns, high cost of data collection, challenges posed by children's limited attention span and inconsistent adherence to prompts during reading tasks.

In response to the challenges associated with collecting real children's speech data, an alternative strategy has emerged, involving the generation of synthetic datasets using TTS models. TTS offers a solution to bypass the difficulties involved in collecting and annotating real children's speech data. While some studies have explored the application of TTS for adult ASR, either through the direct use of synthetic speech for training or as a form of data augmentation [134, 259], synthesising children's speech introduces a unique set of challenges. The inherent substandard and imprecise pronunciation in children's speech [135] represent an obstacle, raising concerns that the direct use of synthetic data may lead to a decrease in performance [135, 260].

In this chapter, we introduce a novel technique known as *Double-Way Adapter Tunning (DWAT)*, to enhance children ASR models, even in scenarios where imperfect data is employed as augmentation. Our approach involves the separate use of additional Adapter modules for synthetic data during the fine-tuning process. As demonstrated in Chapter 5, the effectiveness of Adapters in bridging the gap between a source and target domain, while retaining knowledge from pre-trained models for children's speech, has been established. Therefore, we hypothesise that Adapters can be extended to mitigate the domain mismatch between real and synthetic data in the context of children's speech. To accomplish this, we propose our DWAT approach, a two-step training procedure, drawing inspiration from the methodology introduced in [171].

In the initial step, Adapters are exclusively trained using synthetic data, while keeping the pre-trained model frozen. This phase allows the Adapters to specialise in addressing the nuances introduced by synthetic data, including its imperfections. In the subsequent step, fine-tuning is performed, modifying both the trained Adapters and the entire model weights. During this fine-tuning process, a combination of synthetic and real data is provided to the model. Importantly, our approach introduces a crucial distinction between synthetic and real data throughout the fine-tuning process. Synthetic data passes through the Adapter layers, enabling the Adapters to handle synthetic characteristics while still contributing to the full model tuning. In contrast, real data bypasses these Adapters as it does not exhibit the imperfections inherent in synthetic data. We hypothesise that this meticulous differentiation could enable the effective use of imperfect synthetic data, ultimately leading to an enhancement in ASR performances.

In this chapter, our objective is to address the following research questions: *Is it possible to use children's synthetic speech to extend the amount of children's data? How can we control the quality and speakers' variability?*

6.2 Enhancing ASR Performance through TTS Data Augmentation

The evolution of TTS systems, achieving human-like quality, presents a promising avenue for effective TTS-based data augmentation in ASR. As exemplified in studies such as [134], this approach entails the generation of synthetic speech from text using TTS models and incorporating it with real speech during the training process, leading to notable performance improvements. Importantly, this data augmentation approach is not limited to well-resourced tasks and has demonstrated success even in low-resource scenarios, as illustrated in [261]. However, despite its efficacy, TTS-based data augmentation offers only modest improvements, primarily owing to the persistent challenge of domain mismatch between synthetic and real speech. Indeed, even with human-like qualities, some generated utterances suffer from artefacts or wrong modelisation.

To address the disparity between synthetic and real data, an approach based on shared intermediate representations has been proposed in [262]. Instead of using Mel-scale filterbanks, this method employs discrete intermediate representations, obtained with a Vector-Quantised Wav2vec [263], that are shared by both the TTS and ASR systems. This approach has demonstrated promising results, emphasising its potential in bridging the gap between synthetic and real data in the context of ASR. However, it is essential to note that implementing such a strategy necessitates training both the ASR and TTS systems from scratch. This requirement may pose challenges or may not be feasible, particularly in the context of low-resource ASR scenarios.

An alternative approach to address the domain mismatch between synthetic and real speech relies on data selection techniques, as proposed by [135]. This approach focuses on selectively choosing high-quality synthetic speech data to mitigate the challenges associated with imperfect data augmentation. The data selection process ensures that only the most reliable and accurate synthetic speech samples are incorporated during the augmentation process. One advantageous aspect of this work is that it does not necessitate more complex training for the TTS or ASR models. Instead, it operates as an off-the-shelf selection on top of the TTS engine. This characteristic underscores the practicality and ease of its integration into existing TTS systems. In the study conducted by Wang et al. [135], the effectiveness of using i-vector speaker-embedding cosine similarity between reference and generated utterances as a metric for data selection was demonstrated. This metric outperformed other measures, including error rate, acoustic posterior, and synthetic discriminator.

More recently, [260] introduced the Synth++ framework, which employs a similar data selection approach, called rejection sampling. This rejection sampling method relies on the output of a DNN, which is trained on a 5-dimensional features vector derived from a pre-trained ASR model. The goal of this DNN is to discriminate whether the data is real or synthetic. To this end, the 5-dimensional features vector for each utterance encompasses CE loss, CTC loss, WER, lengths of tokens in the predicted text and the length of tokens in the target text. In addition to rejection sampling, Synth++ introduces the use of separate batch normalisation for real and synthetic data. During the training process, when synthetic data is fed into the model, it undergoes distinct normalisation layers. The incorporation of this separated normalisation has been demonstrated to significantly reduce the mismatch between real and synthetic data during training, leading to notable improvements in WER scores.

6.3 Closing the synthetic and real mismatch gap with Adapters: The Double Way Adapter Tunning approach

The literature extensively documents the effectiveness of Adapters in both speech and NLP tasks [246, 264, 265]. Additionally, Chapter 5 further reinforces the positive results, demonstrating the efficacy of Adapter modules in children's ASR. These findings highlight the role Adapters can play in addressing the mismatch between a source and target task, showcasing their ability to capture task-relevant information while preserving valuable knowledge of a frozen pre-trained model.

In alignment with these findings, [171] introduced the Domain Responsible Adaptation and Fine-Tuning framework (DRAFT) framework for children's ASR. The authors aimed to reduce the mismatch between adult and children speech data in SSL models. To this end, Adapters were integrated and trained within the SSL model, followed by an additional adaptation phase where both the Adapters and the SSL model were fine-tuned. The DRAFT framework proved to be effective, resulting in improved performances.

Inspired by the successes of the DRAFT framework and Synth++, our DWAT approach integrates Adapter modules as a replacement for the separate normalisation layers in the Synth++ framework. Additionally, we incorporate the multiple-step adaptation process from the DRAFT framework. Finally,

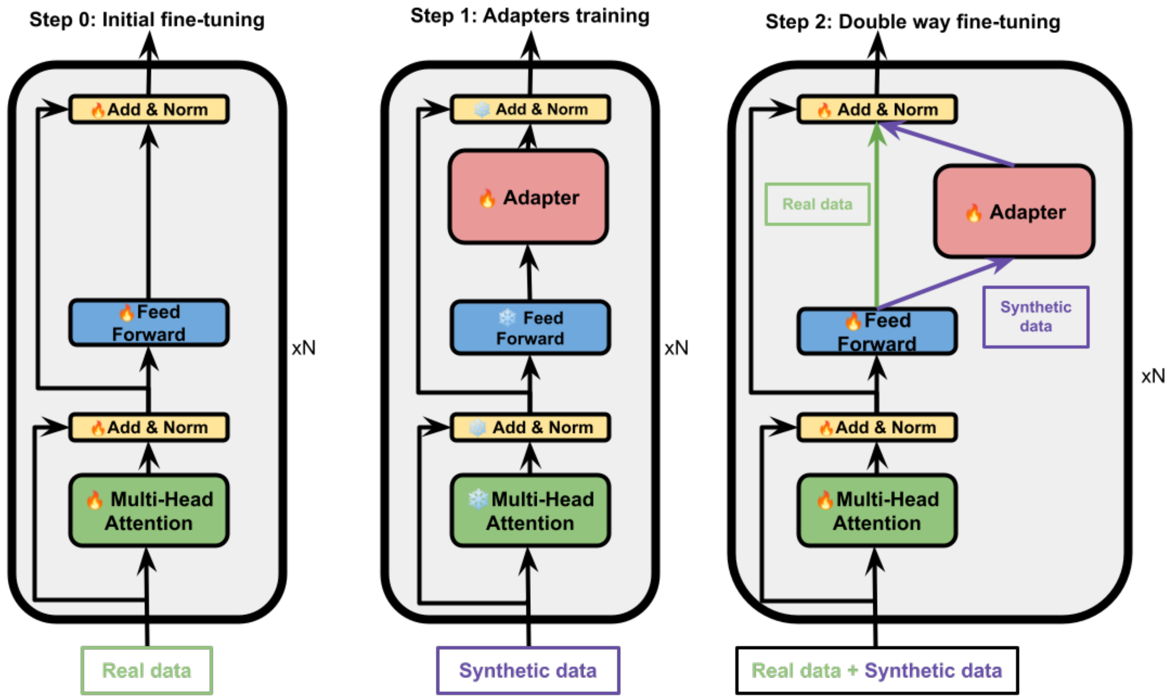


Figure 6.1: Overview of "Double way Adapter fine-tuning" within the context of a Transformer model

we used filtered synthetic data, implementing a speaker-embedding cosine similarity metric to retain synthetic utterances that exhibit high-quality generation, as proposed by [135]. This combination constitutes a novel strategy to tackle the domain mismatch between real and synthetic children's speech for ASR.

In our DWAT framework, as illustrated in Figure 6.1, we introduce two additional steps following the standard ASR model training with children's data (Step 0). Step 1 involves the standard training of Adapter modules. While keeping the pre-trained ASR model parameters fixed, the Adapter weights are trained on the target data, here TTS utterances. These Adapter modules are strategically placed after the Transformer FFN component. The goal is to learn a projection that aligns synthetic children's speech with real children's speech within each Transformer layer. This approach allows the model to retain knowledge about children's speech, while the Adapters aim to capture the inherent synthetic characteristics of the different TTS utterances. This step is crucial, as Adapter modules require this learning process to efficiently bridge synthetic and real speech. Without this Adapter training step, the subsequent fine-tuning in Step 2 could be more challenging and less effective, as the bridge learned by Adapters has to be established concurrently with the normal fine-tuning process.

In Step 2, we fine-tune both the Adapters trained in Step 1 and the entire pre-trained ASR model using a mix of synthetic and real data. A pivotal aspect of our approach lies in how we handle data flow within the model. Real samples bypass the Adapter modules as they do not need further adjustments, directly passing through the original ASR model components. In contrast, synthetic data goes through the Adapters for necessary modifications to better align with real children's speech characteristics. This differentiated treatment of data used Adapters to address the mismatch of the synthetic data, allowing the rest of the ASR model to focus on learning normal children's characteristics. This approach has the potential to enhance the overall performance of the ASR system.

During the inference phase, the Adapter modules become unnecessary and are discarded since the test data only contains real samples. It is essential to note that Steps 1 and 2 can be iteratively repeated with newly generated synthetic data. However, this aspect is not investigated in this thesis and serves as a subject for future research.

In summary, our proposed DWAT approach uses Adapter modules to enhance the performance of a pre-trained ASR model through the incorporation of synthetic data augmentation. This innovative methodology introduces a two-step process involving the training of Adapter and subsequent fine-tuning of the ASR model with a mix of synthetic and real data using Adapters to bridge the domain gap between real and synthetic children's speech.

6.4 Overview of the automatic speech recognition and text-to-speech systems

6.4.1 Transformer architecture for ASR

For this experiment, we used a pre-trained Transformer model[1] based on the SpeechBrain toolkit [241] as the ASR component. The choice of the Transformer architecture was motivated by the positive results obtained for children's ASR in previous chapters. We will not delve into the details of the Transformer architecture here, as a comprehensive explanation was provided in Section 4.2. For this experiment, the model was trained on the LibriSpeech dataset [31], comprising 12 Encoder Transformer layers and 6 Decoder Transformer layers, each with a dimension of 512. For training, we employed a combination of Seq2Seq and CTC loss functions, with respective weights of 0.7 and 0.3. Additionally, we integrated a Transformer language model trained on a 10 million-word corpus of transcriptions from the LibriSpeech dataset. It is essential to note that this particular instantiation of the Transformer model is different from the one employed in the previous chapters of our study.

6.4.2 Multi-speaker text-to-speech: YourTTS

In this work for the TTS component, we used the pre-trained YourTTS model[3] proposed by [266] based on the Coqui toolkit. YourTTS is a zero-shot multi-speaker and multilingual TTS system that is built upon the VITS model. It incorporates several novel modifications to enable zero-shot multi-speaker and

[1]https://huggingface.co/speechbrain/asr-transformer-transformerlm-librispeech/tree/916c9ff
[3]https://coqui.ai/blog/tts/yourtts-zero-shot-text-synthesis-low-resource-languages

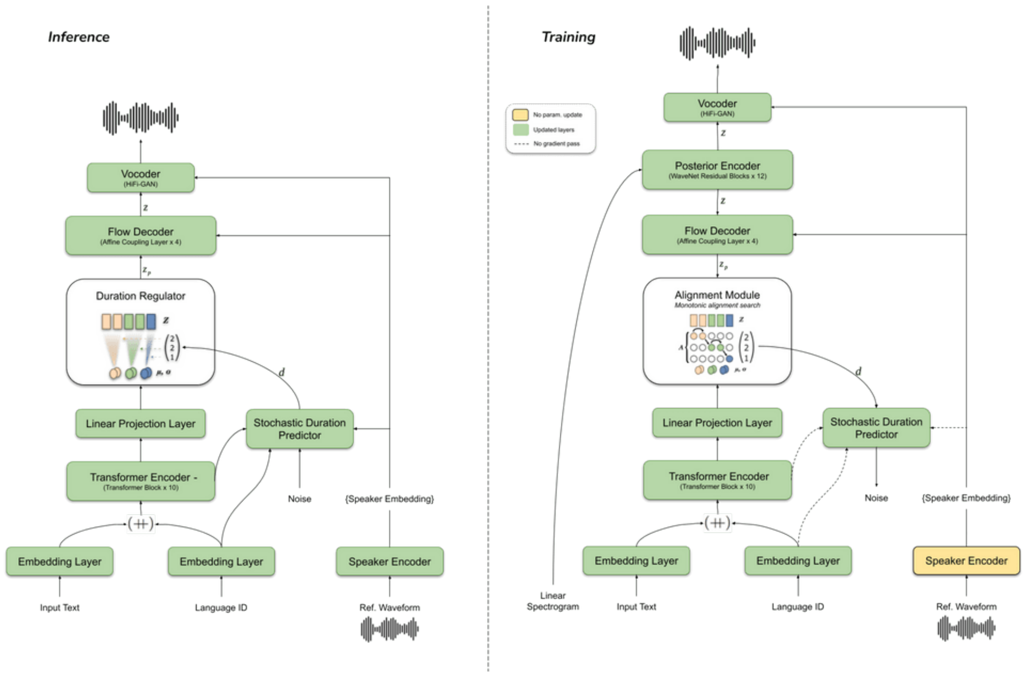


Figure 6.2: Architecture of the YourTTS model[2]

multilingual synthesis. An overview of the YourTTS architecture is presented in Figure 6.2. YourTTS features a 10-layer Transformer-based text Encoder with 196 hidden channels. This Encoder, adaptable for multilingual use, employs a 4-dimensional language embedding concatenated with the embedding of each input character. However, for the purpose of our experiment, we exclusively use the English language. The Decoder comprises four affine coupling layers [267], each incorporating four WaveNet blocks [268] to ensure high-quality speech generation. The model also uses the HifiGAN vocoder [269]. Notably, YourTTS adopts an end-to-end approach, connecting the vocoder to the TTS model through a Variational AutoEncoder (VAE) [270].

To enhance its capabilities as a multi-speaker and zero-shot TTS system, YourTTS integrates a speaker Encoder, specifically the Half / Attentive Statistics Pooling (H/ASP) speaker Encoder [271], generating 512-dimensional speaker embeddings for each utterances. This speaker Encoder model is compromising a CNN layers, followed by four Resnet layers [122], an attentive statistic pooling and an output linear layer. These embeddings serve as reference speakers for the TTS model, enabling zero-shot multi-speaker capabilities. To this end, the authors conditioned all affine coupling layers of the flow-based Decoder, the posterior Encoder, and the vocoder on external speaker embeddings. Additionally, a Speaker Consistency Loss (SCL) is added to the final loss to further enhance the multi-speaker ability of the model. The SCL is formally expressed as follows:

[3]Figure taken from https://coqui.ai/blog/tts/yourtts-zero-shot-text-synthesis-low-resource-languages

$$L_{SCL} = \frac{-\alpha}{n} \cdot \sum_{i}^{n} cos\_sim(\phi(g_i), \phi(h_i)) \tag{6.1}$$

where let $\phi(\cdot)$ be the function outputting the speaker embeddings, cos_sim is the cosine similarity function, α is a positive real number controlling the influence of the SCL in the final loss, n is the batch size and g and h represent, respectively, the ground truth and the generated speaker audio.

The YourTTS model was trained using three languages: English with VCTK [272], Brazilian Portuguese with TTS-Portuguese Corpus [273], and French with the French set of the M-AILABS dataset [274]. This training corpus totalled 229 hours of speech data and involved 115 speakers. For a more in-depth understanding of the YourTTS architecture and training process, detailed information can be found in the original paper [266].

6.5 Experimental setup

This experiment uses the My Science Tutor (MyST) Children Speech Corpus, as the "Real" children'speech set. We used the Myst dataset here as it was employed in prior experiments conducted on this dataset within the context of this thesis, as detailed in Table 4.1. For a more comprehensive understanding of the MyST corpora, additional details are provided in Section 2.4.

6.5.1 Synthetic data

To address the potential performance gap caused by the YourTTS model being trained solely on adult data and never exposed to children's data, we initiated the process by fine-tuning the YourTTS model using the MyST training set. In this study, two TTS systems were developed, each with distinct parameter settings, to examine their respective performances and output quality under varying conditions. The two TTS systems, namely TTS_1 and TTS_2, were developed with distinct parameter settings. The first model, TTS_1, underwent fine-tuning for 250 epochs, without incorporating the speaker Encoder SCL loss. In contrast, the second system, TTS_2, was fine-tuned for 50 epochs, incorporating the speaker Encoder SCL loss. This incorporation should improve the alignment between the generated speech and the reference speaker embedding provided to the model. These variations in training strategies enable a thorough examination of the model's performance and output quality under varying conditions,

The first TTS model, TTS_1, was used to generate 300 hours of synthetic data referred to as $\textit{Synth}_1$. The second TTS model, TTS_2, was employed to generate a larger volume of synthetic data, up to 1,000 hours, denoted as *Large* $\textit{Synth}_2$. To compare the performance of TTS_1 and TTS_2, a subset of 300 hours was extracted from the *Large* $\textit{Synth}_2$ dataset, called $\textit{Synth}_2$. The full 1,000-hour set was exclusively used to evaluate the impact of different amounts of synthetic data on the performances of our DWAT approach.

The speech synthesis process using the YourTTS model requires a text transcription and a speaker-embedding vector to generate a synthetic utterance. Therefore, to generate an utterance in both $Synth_1$ and *Large* $Synth_2$, we randomly selected two utterances from the Myst training set, one designated for extracting the speaker embedding and the other exclusively used for its text transcription. The Myst training set comprises 60,897 utterances, therefore resulting in a vast pool of potential combinations, totaling 3,708,444,609 unique combinations. This extensive range of possibilities was strategically harnessed to introduce a deliberate mismatch between the selected speaker embeddings and their associated transcriptions. This intentionally large range of possibilities served as a mechanism to infuse novel variability into the synthetic data, thereby introducing characteristics not present in the original real corpus.

The decision to use MyST transcriptions for generating synthetic data was driven by the aim to expose the TTS model to the unique transcription style present in the MyST dataset. This style encompasses elements such as "UM" hesitations. By adopting this strategy, we intended to enhance the model's ability to learn and reproduce the specific transcription characteristics of the MyST data.

To ensure the quality of the synthetic utterances in $Synth_1$ and *Large* $Synth_2$, we extended the approach proposed by [135], which involves using cosine similarity between the speaker embeddings of the reference and synthetic utterances as a data selection criterion. Instead of using i-vector speaker embeddings, we opted for an x-vector approach and employed a different speaker embedding extractor than the one used by the YourTTS model (to prevent conflicts with the SCL loss). Specifically, we used a pre-trained ECAPA-TDNN [275] based x-vector extractor trained on VoxCeleb[4]. To this end, during the generation process, we applied a cosine similarity threshold of 0.75 to discard all poorly generated synthetic utterances. While exploring data selection mechanisms, we considered the rejection sampling method suggested by Synth++ [260] but found it unsatisfactory, ultimately opting for speaker-embedding similarity. To comprehensively evaluate the impact of this data selection, we also created *Unfiltered* $Synth_1$ and *Unfiltered* $Synth_2$, two 300-hour corpora of synthetic data generated without using speaker embedding data selection, created with TTS_1 and TTS_2, respectively.

6.5.2 Experiments

The evaluation of our DWAT approach involved a comprehensive series of experiments, systematically comparing its performance with existing methods. The initial phase focused on the development of baseline models, by fine-tuning a pre-trained adult model to children's speech. This fine-tuning used exclusively the *Real* data for 20 and 25 epochs, corresponding to step 0 in Figure 6.1.

Subsequent assessments involved the exploration of TTS data augmentation by combining $Synth_1$ and $Synth_2$ data with the *Real* data. To evaluate the impact of data filtering, we assess the results of filtered TTS data augmented with their unfiltered counterparts, denoted as *Unfiltered* $Synth_1$ and

[4]https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

Method	TTS_1	TTS_2
Real (20 epochs)	12.99%	
Real (25 epochs)	13.15%	
Real + Unfiltered *Synth*	13.41%	13.24%
Real + *Synth* [135]	13.09%	12.98%
Synth alone	40.58%	40.21%
Two step adaptation	13.49%	13.46%
Norm double-way (from adult)	12.89%	13.04%
Norm double-way (from children)	13.56%	13.87%
DWAT (Ours)	**12.42%**	**12.31%**

Table 6.1: Results of the different use of synthetic data augmentation approaches in WER. Collomn TTS_1 and TTS_2 correspond to the respective model used to generate the synthetic data.

Unfiltered Synth$_2$. Additionally, we analyse the performance of models trained solely on the filtered synthetic data, aiming to discern the presence of a mismatch between the *Real* and synthetic domains. In addition, we conducted an evaluation of a two-step adaptation process. Initially, the model is fine-tuning with synthetic data only, followed by a subsequent fine-tuning phase exclusively with the *Real* data. All models involved in these data augmentation experiments underwent training for 20 epochs.

Moreover, our exploration delved into the incorporation of double-way normalisation, where distinct normalisation layers within the Transformer architecture were applied to synthetic data during the fine-tuning process. This concept was inspired by the methodology introduced in Synth++ [260]. In one scenario, fine-tuning was conducted directly from the adult model for 20 epochs, incorporating both filtered synthetic and *Real* data. This specific configuration is denoted as *Norm double-way from adult*. To facilitate a comprehensive comparison with our DWAT approach, we introduced an alternative scenario where the double-way normalisation model was trained for 5 epochs, using the baseline model as a starting point for the fine-tuning. This particular configuration is referenced as *Norm double-way from children*.

Finally, we implemented our *DWAT* approach, training the models for 5 epochs with the baseline model as initialisation. In section, 6.6, we will delve into the exploration of various hyper-parameter configuration's influence on the DWAT framework.

6.6 Results and discussion

6.6.1 Comparison with existing approaches

Table 6.1 presents the results of the various approaches. The baseline models, achieved by fine-tuning an adult model on children's speech using *Real* data, yielded a WER score of 12.99%. However, extending the training to 25 epochs led to overfitting, resulting in a subsequent reduction in performance with a WER of 13.15%.

Moving to the impact of synthetic data augmentation, incorporating filtered synthetic data (*Real +*

Amount of TTS data	WER ↓
0h	12.99%
10h	12.73%
50h	12.54%
100h	12.49%
300h	**12.31%**
600h	12.57%
1000h	13.14%

Table 6.2: Results of the different number of hours influence in our DWAT approach with *Large Synth*$_2$ data

Synth) as opposed to the unfiltered counterpart (*Real + Unfiltered Synth*), revealed a notable 2% relative WER improvement when using the filtered versions. However, relying solely on filtered TTS speech (*Synth* alone) without any *Real* data resulted in a was found to be ineffective with around 40% WER on the *Real* test set for both TTS$_1$ and TTS$_2$. This discrepancy underscores the considerable domain mismatch between real and synthetic data, emphasising the imperative need for further approaches to mitigate this gap in the context of synthetic data augmentation.

Furthermore, we explored the two-step adaptation process, wherein the adult model was initially fine-tuned using only filtered *Synth* data, followed by subsequent fine-tuning with only *Real* speech. However, the results indicated a slight decrease in performance in both cases compared to the baseline model, with WER scores of 13.49% and 13.46%, for respectively TTS$_1$ and TTS$_2$. This decline could be attributed to the potential gap between the characteristics of TTS data, which may be more pronounced than the gap between adult and children data.

During our experiments, the use of double batch normalisation proved to be ineffective in enhancing the performance of the baseline model. On the contrary, it resulted in a 5% relative decrease in WER when evaluated with the baseline model as initialisation (from children). It is worth noting that, when training from the adult model, the results remained consistent with those of the baseline models, showcasing no observed improvement. These outcomes underscore the limitations of double batch normalisation in addressing the domain mismatch between real and synthetic speech data within a Transformer model. This emphasises the significance of exploring alternative strategies to effectively bridge this gap.

Our proposed DWAT approach, using the baseline model as the initial step (step 0), emerged as the most effective among all the methods examined. It demonstrated a notable 4% and 5% relative improvement in WER over the baseline when evaluated on *Synth*$_1$ and *Synth*$_2$ respectively. Additionally, we observed that our approach, while training 5 additional epochs, did not lead to overfitting, especially compared to longer training on the *Real* set (*Real 25 epoch*), showcasing its potential in mitigating the challenges posed by domain mismatch in the context of synthetic data augmentation for children ASR.

Location	Bottleneck size	5 epochs	20 epochs
Encoder	64	12.58%	12.24%
Encoder	128	12.31%	12.45%
Encoder	256	**12.25%**	12.32%
Encoder	1024	12.42%	**12.22%**
Encoder	2048	12.57%	12.47%
Encoder-Decoder	128	12.45%	12.48%
Skip step 0	256	12.30%	-
Skip step 0 and 1	256	13.28%	-

Table 6.3: Results of the different configurations of Adapter double-way approach on 300h of $Synth_2$

6.6.2 Influence of synthetic number of hours

Given the potential of our approach to generating a theoretically "infinite" amount of TTS data, we aimed to evaluate the impact of varying synthetic data quantities on the efficacy of the DWAT approach. To this end, we assess the influence of different quantities of synthetic data taken from *Large* $Synth_2$, as detailed in Table 6.2. The findings from our investigation reveal that utilising a small quantity of synthesised speech, ranging from 10 to 50 hours, results in limited improvements in ASR performance. Conversely, an excessive volume of TTS data, spanning from 600 to 1,000 hours, has the potential to introduce undesirable noise into the system. Notably, only the configuration with 1,000 hours failed to improve the baseline score of 12.99%. Consequently, achieving an optimal equilibrium becomes crucial. In our experiment, we found that this equilibrium of synthetic speech fell within the same range as the number of hours available for real speech, around 100 to 300 hours. This balance aims to enhance the robustness of ASR performance with new variabilities while concurrently preventing the introduction of excessive noise, thereby maximising the effectiveness of the DWAT approach.

6.6.3 Impact of DWAT different hyperparameter configurations

To further evaluate the robustness of our approach, we conducted experiments with various hyperparameter configurations of DWAT. This extensive analysis involved exploring different Adapters's bottleneck sizes, ranging from 64 to 2048, varying the number of final fine-tuning training epochs (5 and 20), investigating the integration of Adapters in the Decoder of the transformer model, and performing an ablation study by skipping step 0 and both steps 0 and 1 in the DWAT process.

The results of the different configurations are summarised in Table 6.3. These results provide insights into the optimal configuration, revealing that the most effective setup consists of Adapters with a hidden dimension size of 1024, placed in the Encoder only, coupled with 20 training epochs for the final fine-tuning. This configuration resulted in a remarkable 6% relative WER reduction compared to the baseline. Notably, all configurations demonstrated superior performance to the baseline, underscoring the overall effectiveness of our DWAT approach.

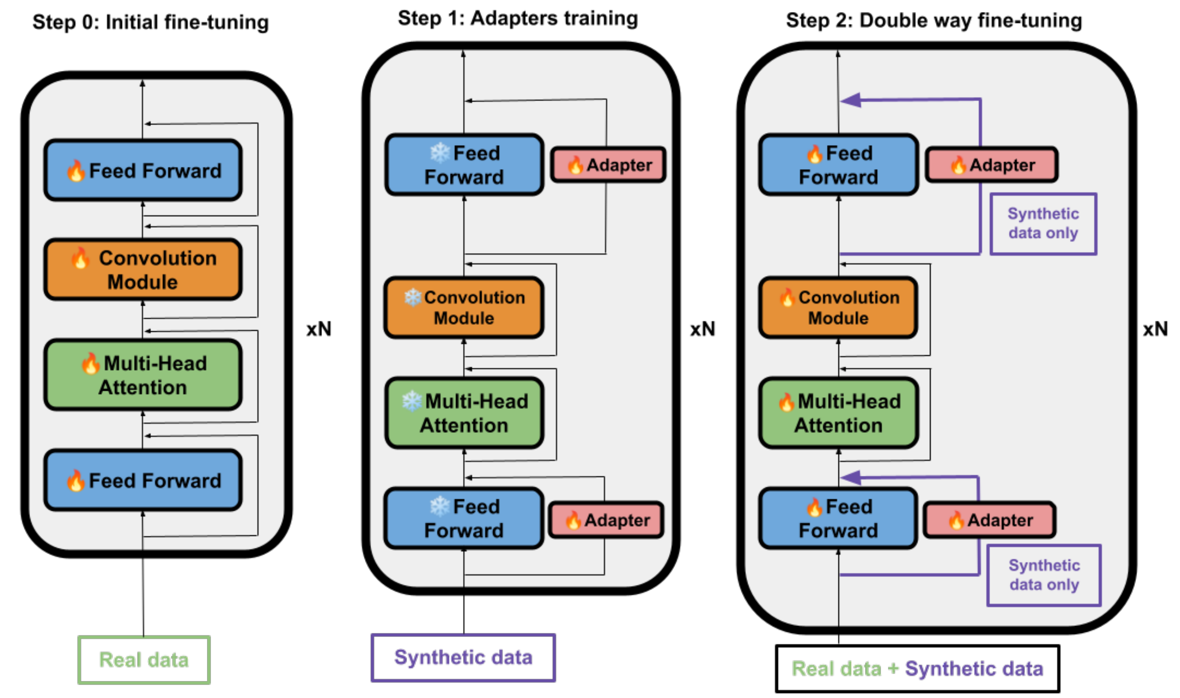


Figure 6.3: Overview of "Double way Adapter fine-tuning" within the context of a Conformer architecture

Furthermore, our results indicate that extending the training period in was particularly advantageous for larger Adapter bottleneck sizes, demonstrating no signs of overfitting. Moreover, the integration of Adapters into the Decoder did not lead to a significant improvement in results. This outcome can be attributed to the higher acoustic variability typically captured in the Encoder, coupled with the fact that the same transcriptions from the Myst dataset were used for synthetic speech utterances, thereby diminishing the potential for further improvement in the Decoder. Lastly, skipping step 0 (initial children pre-training) did not result in a significant degradation in performance. However, when both step 0 and step 1 (initial pre-training and Adapter pre-training) were skipped, performance degradation occurred, underscoring the critical role of the Adapter pre-training phase (step 1) in achieving good performances.

6.6.4 Extension DWAT to the Conformer architecture

Building upon the positive outcomes observed in previous chapters, where the Conformer architecture showcased superior performance compared to the regular Transformer in children's ASR, and considering the efficacy of Adapters in the Conformer architecture, we decided to assess the performance of our DWAT within the Conformer architecture. Additionally, since Synth++ [260] was initially designed for the Conformer architecture, we opted to evaluate the Synth++ double-normalisation layers compared to our DWAT in this setting.

Due to the demonstrated effectiveness of the TPA configuration for the Conformer model, we chose to implement our DWAT framework within the Conformer architecture using two Adapters per layer, specifically, one for each FFN module—in parallel. The integration of our DWAT framework within the Conformer using TPA is illustrated in Figure 6.3.

In this experiment, we employed the same pre-trained adult Conformer model as employed in the preceding chapters. This model consists of 12 layers of Conformer in the Encoder, followed by 6 Transformer layers in the Decoder, all with a hidden dimension of 512. The same language model was used and aligns with the experiments conducted in the Transformer model detailed in the preceding sections. In light of the insights derived from the exploration of hyperparameters in Section 6.6.3, we chose Adapters with a hidden dimension size of 512, solely in the Encoder and using the TPA configuration. Regarding training, both Step 0 and Step 1 were trained for 30 epochs, while Step 2 was only 10 epochs.

Method	WER Score ↓
Adult Frozen model	21.75%
Real only	12.28%
Unfiltered $Synth_2$	37.85%
Unfiltered $Synth_2$ + *Real*	12.30%
$Synth_2$ alone	31.72%
$Synth_2$+ *Real*	12.02%
Two-step adaptation	12.51%
Synth++ (Norm double way)	11.80%
DWAT	**11.64%**

Table 6.4: Scores for the different methods of synthetic data augmentation within the Conformer architecture

The results obtained from the extension of DWAT to the Conformer architecture are presented in Table 6.4. The frozen adult model achieved a WER score of 21.75%. When considering only real data (*Real* only), the WER score improved significantly to 12.28%. Which is already outperforming the best results of the Transformer.

When using unfiltered synthetic data alone (*Unfiltered* $Synth_2$) method, it resulted in a higher WER score of 37.85%, showcasing, once again, the mismatch between synthetic and real data. However, when combined with real data (*Unfiltered* $Synth_2$ + *Real*), the results are closer to the real baseline with a WER score of 12.30%. Similarly, using the filtered $Synth_2$ alone resulted in a WER score of 31.72%, but when combined with real data ($Synth_2$ + *Real*), the WER score decreased significantly to 12.02%. This differs from the results observed in the Transformer architecture, as in the Conformer scenario, data filtering has been able to improve the WER score compared to the baseline. Additionally, similarly to the Transformer, the two-step adaptation method yielded a WER score of 12.51%, falling behind the baseline score.

Further insights into the performance of the more sophisticated methods are provided in the context of the Conformer architecture. Specifically, the Synth++ method, which incorporates separated batch normalisation within each Convolution module for TTS data, achieved a WER score of 11.80%. This contrasts with the results observed in the Transformer experiment, highlighting the efficiency of this approach specifically within Conformer models. On the other hand, the DWAT method consistently outperformed all other methods, showcasing a remarkable WER score of 11.64%. These findings emphasise the effectiveness of both the Synth++ and DWAT methods in enhancing the Conformer architecture's performance on the given task.

However, it is important to highlight that while both Synth++ and DWAT demonstrate efficacy within the Conformer setup, the DWAT approach consistently outperforms other methods in both Transformer and Conformer configurations. This consistent superiority underscores the relevance and robustness of the DWAT approach in improving the overall performance of the children ASR models across different architectures.

In summary, the experimental results demonstrate the effectiveness of leveraging synthetic data, real data, and advanced adaptation methods within the Conformer architecture. The DWAT method, in particular, stands out as the most successful approach in minimising the WER.

6.7 Summary and discussion

In this chapter, we aimed to answer to the following research questions: *Is it possible to use children's synthetic speech to extend the amount of children's data? How can we control the quality and speakers' variability?*

We provide positive responses to both research questions through the introduction of our novel methodology: the Double Way Adapter Transfer procedure, which combines Adapters and synthetic data augmentation for children's speech recognition. Our two-phase training strategy consists of the initial training of Adapter modules using synthetic data, followed by the fine-tuning of Adapters and the entire model weights. Importantly, during this fine-tuning process, real data bypasses the Adapter modules. This distinctive dual-pathway approach resulted in notable improvement over baseline and previous techniques across various configurations. Importantly, our approach demonstrated robustness across different ASR architectures, TTS model fine-tuning parameters, Adapter sizes, number of epochs, and varying amounts of synthetic data. Furthermore, our study showcases the controllability of TTS output quality without necessitating direct modifications of the TTS model. This filtering process, performed before applying the DWAT, is achieved through the use of pre-trained x-vectors speaker embeddings and cosine similarity metrics between reference and generated utterances. This approach, improved from previous work where only i-vectors were used [135], allows the use of synthetic data which align more closely with desired characteristics of real children's speech, contributing to a more controlled augmentation process.

The promising performances observed with the DWAT pave the way for future research avenues. One potential avenue involves adopting an iterative methodology, dynamically incorporating newly generated TTS data while adjusting the proportion of real data and repeating the operation as needed. This iterative fine-tuning approach could potentially provide further insights into the interaction between synthetic and real data, offering opportunities to refine and optimise the training process for improved children's ASR models. Another avenue of exploration could be to extend the DWAT framework to domains beyond the synthetic and real speech data. Investigating the adaptability and effectiveness of DWAT in diverse domains, such as adult and children's speech domains could be an interesting first step towards this direction. Furthermore, considering the rapid evolution of the PETL approaches, the exploration of new modules as replacements for Adapters represents a new research direction. An initial step in this direction will be explored in the upcoming chapter, where we conduct a thorough evaluation of different PETL alternatives. The goal is to identify novel strategies for enhancing children's ASR systems.

7

Alternative approaches to parameter-efficient transfer learning

Contents

7.1 Introduction

In the preceding chapters, we demonstrated the effectiveness of Adapters, particularly residual Adapters, across various tasks associated with children's ASR. These tasks encompassed PETL for children's ASR and mitigating domain mismatch for TTS data augmentation. The success observed in these experiments has prompted a more in-depth analysis of potential alternatives to Adapters. Additionally motivated by the literature, which offers alternative approaches that often yield superior performances and provide more parameter-efficient solutions for other tasks. Indeed, the growing attention towards PETL from the research community has led to the emergence of an array of novel architectures and methodologies. These alternatives extend beyond the conventional Adapters. These innovative approaches have been extensively explored in various domains such as NLP and image processing but remain relatively unexplored in the domains of speech processing and even more for children's ASR.

The objective of this section is to analyse these emerging PETL methods specifically in the context of children's ASR. Drawing from our prior findings, which underscored the significance of fine-tuning FFN modules, we have curated a selection of PETL approaches designed to be integrated with FFN modules. Notably, we excluded PETL approaches centred around the MHSA modules, such as LORA [276], as well as any prompt-related PETL strategies like prompt tuning [277] or prefix tunning [278].

This chapter starts with an in-depth presentation of the selected PETL, elucidating their underlying principles, architectural intricacies, and key characteristics compared to the traditional Adapter in the context of children's ASR. The goal of this chapter is to answer the following research question: *Can we further improve the parameter-efficiency with other architectures?* The evaluation mainly encompasses two dimensions, the performances compared to the full-finetuning and the parameter-efficiency of the method. By systematically benchmarking these alternative PETL methodologies against the conventional Adapter models, our objective is to not only understand the individual strengths and limitations of these approaches but also to evaluate their relative effectiveness in tackling the specific challenges presented by children's ASR.

7.2 Exploring PETL literature alternatives

7.2.1 Scaled Adapters

Scaled or Gated Adapters extend the conventional residual Adapters by introducing a scaling mechanism to the output of the Adapter modules. The concept of scaled-Adapters was initially introduced by [256] and is formally expressed as:

$$adapter(x) = x + s \cdot (W_{up}(f(W_{down}g(x) + b_{down})) + b_{up}) \tag{7.1}$$

Here, $s \in \mathbb{R}$ is a tunable scalar hyperparameter. Notably, some research has proposed to learn directly this scalar value during training as a gate mechanism [265]. The intuition behind this approach is to allow the network to gradually learn to assign weights to the target domain, achieving more fine-grained control of the Adapter activation or deactivation. The scaled learning process should facilitate the regulation of contributions from the Adapters from different layers, enabling the model to adapt and refine its responses based on the specific characteristics of the data encountered during training in a layer-wise fashion. Within the context of our experiments, we decided to use a trainable scalar associated with each Adapter module using a TPA configuration in the Encoder only. These scaling parameters and all Adapter modules were optimised through 30 epochs of training using a learning rate of 8×10^{-4}.

7.2.2 Convolution based Adapters

CNNs have been widely recognised for their effectiveness in exploiting local information, particularly for computer vision tasks [279]. These networks learn shared kernels based on position within localised windows, giving them the ability to capture features such as edges and shapes. This characteristic has also been demonstrated in the field of speech-related tasks, as demonstrated by the success of the Conformer architecture [225].

As a result of this success, CNNs have been naturally incorporated into Adapter modules. This strategic fusion enables Adapters to leverage the spatial processing capabilities inherent in convolutional modules, thereby enhancing their capacity to capture and adapt to different patterns present in the data. The initial approach involved using CNNs as feature extractors within an Adapter, in conjunction with a linear transformation, as detailed by [280], we denoted this approach as Conv-Adapter. We experimented with two versions of Conv-Adapters, where a 1-dimensional CNN is used as a replacement of either the first or second linear of the traditional Adapters, respectively $\textit{Conv-Adapter}_{down}$ and $\textit{Conv-Adapter}_{up}$. More formally, the $\textit{Conv-Adapter}_{down}$ can expressed as followed:

$$\text{Conv-adapter}_{Down}(x) = x + (W_{up}(\text{CNN}_{1D}(x)) + b_{up}) \tag{7.2}$$

While $\textit{Conv-Adapter}_{up}$ is defined as:

$$\text{Conv-adapter}_{up}(x) = x + \text{CNN}_{1D}(W_{down}(x) + b_{down}) \tag{7.3}$$

We also investigated, the [281] Convolutional-Adapter where the CNN is integrated in-between the Up and Down projection of the traditional Adapter, denoted $\textit{Conv-Adapter}_{Middle}$, mathematically expressed as:

$$\text{Conv-adapter}_{Middle}(x) = x + W_{up}(\text{CNN}_{1D}(W_{down}(x) + b_{down}) + b_{up}) \tag{7.4}$$

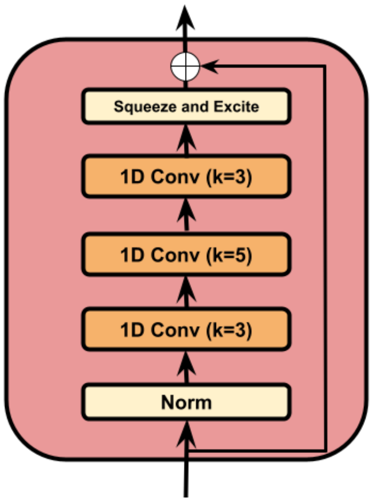


Figure 7.1: The architecture of the ConvPass adapter. k is the kernel size of the 1D convolution. All Convolutions are depth-wise convolutions.

In addition to the Conv-Adapters, an alternative approach involves the use of a fully CNN-based Adapter known as ConvPass, which has demonstrated effectiveness in computer vision tasks [282]. Diverging from conventional Adapters and the previously mentioned Conv-Adapter, ConvPass distinguishes itself by the removal of the up and down linear layers, replaced instead by three CNN layers. In [282], these layers comprise a 1×1 convolution, followed by a 3×3 convolution, and another 1×1 convolution. With Gaussian Error Linear Unit (GELU) activation functions are interposed between these convolutional layers.

For speech-related tasks, a comparable approach was introduced by [283]. The speech-specific ConvPass, illustrated in Figure 7.1, incorporates a layer normalisation layer, followed by three lightweight 1-dimensional CNN layers with kernel sizes of 3, 5, and 3, respectively. Additionally, a squeeze and excite module is integrated into the architecture. The squeeze and excite module, as proposed by [284], facilitates feature recalibration. It consists of a global pooling operation, followed by a linear layer, a ReLU activation, a second linear layer, and concludes with a Sigmoid activation.

During our experiments, all Conv-Adapters and ConvPass configurations were trained for 30 epochs with a learning rate of 8×10^{-4} using a TPA configuration in the Encoder only.

7.2.3 BitFit

Bias-Term Fine-Tuning, known as BitFit, is a PETL method introduced by [285] for NLP tasks. The main idea behind BitFit is to fine-tune only the bias terms and the task-specific classification layer while keeping the rest of the model frozen. This approach aims to achieve efficient fine-tuning with reduced computational requirements in a similar way as our partial fine-tuning experiments of Chapter 4. The fine-tuning of bias terms can be seen as introducing a task-specific shift to the token representations.

The authors highlight three key properties of BitFit. Firstly, it matches a fully fine-tuned model on some NLP tasks, showcasing its ability to maintain comparable results while significantly reducing the number of parameters to be trained. Secondly, BitFit is designed to adapt to tasks arriving sequentially, eliminating the need for simultaneous access to all datasets. This adaptability enhances the model's versatility in handling diverse tasks with varying data distributions over time. Thirdly, BitFit exhibits parameter efficiency by fine-tuning only a small subset of the model's parameters.

In our experiment, the pre-trained model and children's ASR task shared the same output dimensions and character encoding. In consequence, we excluded the fine-tuning of the task-specific classification layer concentrating solely on the tuning of the bias terms. The training process consisted of 30 epochs with a learning rate of 8×10^{-4}.

7.2.4 Scale and Shift features

Scale and Shift Features (SSF) was introduced an PETL alternative approach by Lian and al. [286] for image classification. The primary objective of SSF is to establish a generalised method for efficient model fine-tuning without the introduction of task-specific inference parameters. Drawing inspiration from feature modulation techniques such as [287, 288], the SSF method modulates deep features extracted by a pre-trained model by scaling and shifting them to match the distribution of a target dataset. The intuition behind SSF comes from the inherent disparities in data distributions between upstream and downstream datasets. Directly applying model weights trained on an upstream dataset to a downstream dataset frequently leads to performance degradation due to the disparities between the two datasets [289]. The SSF method addresses this challenge by introducing scale γ and shift β parameters, which could be considered as the variance and mean used to modulate the features extracted from the pre-trained model. This modulation ensures that the adapted features align with the characteristics of the upstream dataset. Formally, given an input x, the modulated output y is calculated by:

$$y = \gamma \odot x + \beta \tag{7.5}$$

Notably, the scale and shift parameters in SSF remain independent of any input and have a unified learnable parameter space for different tasks. Another notable advantage of SSF is that the scale and shift weights can be seamlessly integrated into the original pre-trained weights of the linear layers during model re-parameterization in the inference phase. This integration avoids the need for additional parameters removing the extra-computation time of other PETL such as Adapters.

In practice, the SSF modulations are introduced after each module of the Conformer architecture, namely the FFNs, MHSA and Convolution modules. In the original paper, they also fine-tuned the head

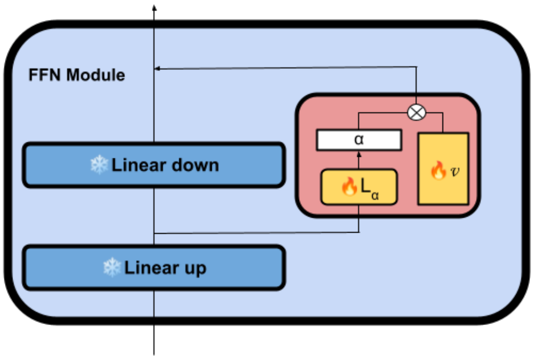


Figure 7.2: AdapterBias, consisting of a linear layer L_α and a vector $\mathcal{V}$, is added after the second feed-forward layer only in each FFN module.

layers. However, as our upstream and downstream tasks, respectively adult and children ASR, share the same output dimension and character encoding, we do not fine-tune this extra layer and we only use the SSF method for each Conformer's modules. Our training compromises 30 epochs with a learning rate of $8 \cdot 10^{-4}$.

7.2.5 AdapterBias

Following the success of BitFit [285], which aims to introduce task-specific shifts to each output representation by selectively fine-tuning only the bias terms of a pre-trained model, recent research has suggested that certain tokens may hold more importance than others for specific tasks. While BitFit uniformly applies the same shift across all tokens regardless of their relevance to the task, [290] proposes AdapterBias to address this limitation.

Compared to traditional residual Adapters, that are used in parallel or in series with the FFN module, AdapterBias is plugged directly inside of the FFN module, added after the second feed-forward layer only. In terms of design, AdapterBias comprises two essential modules: a vector $\mathcal{V}$ and a linear layer L_α. The vector $\mathcal{V}$ represents a task-specific shift added to the output of each FFN module. On the other hand, the linear layer L_α generates a token-dependent weight vector $\alpha = [\alpha_1, \alpha_2, ..., \alpha_m]^T$, where α_i is the weight associated with the shift of the i^{th} token. By applying these token-specific weights to the task-specific representation shift $\mathcal{V}$, AdapterBias focuses on tokens which are more crucial to the task, allowing efficient and fine adaptation to various downstream tasks, as demonstrated in NLP tasks [290].

The output of AdapterBias is defined as the bias B represented as the outer product of $\mathcal{V}$ and the learned weights vector α is summed to the outputs of the second feed-forward layer of the FFN module. Mathematically, the output of AdapterBias B is expressed as follows:

$$B = \mathcal{V} \otimes \alpha^T \tag{7.6}$$

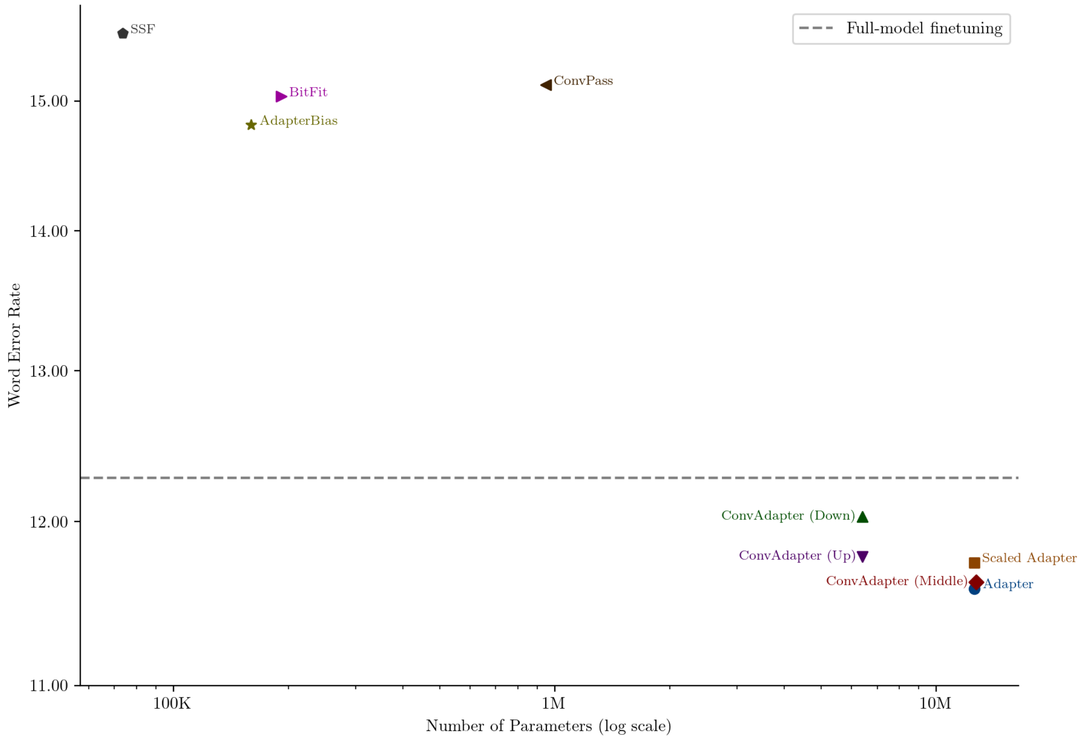


Figure 7.3: Different parameter efficient procedure for children ASR in Conformer model

Here, *otimes* denotes the element-wise multiplication of the task-specific shift vector $\mathcal{V}$ and the token-dependent weight vector α. In our experimental setup, we conducted the training for AdapterBias by integrating them into the second linear layer of each of the two FFN modules within the Conformer architecture. We trained AdapterBias for 30 epochs, employing a learning rate of 8×10^{-4}.

7.2.6 Experimental setup

In all experiments to assess the efficiency of PETL alternatives, we employed the same pre-trained Conformer model outlined in Section 4.4.4. As a training dataset, we used the My Science Tutor (MyST) Children's Speech Corpus. This dataset has been consistently used in prior experiments within the context of this thesis, and further details about the MyST corpora are provided in Section 2.4.

7.2.7 Results of the different PETL methods

The results of various PETL methods are summarised in Table 7.1. Notably, the traditional residual Adapter approach using the TPA configuration emerges as the most effective, achieving an 11.58% WER

PETL Method	WER ↓	Trained Parameters
Adapter	**11.58%**	12.6M
Scaled Adapter	11.74%	12.6M
ConvAdapter(Down)	12.03%	6.4M
ConvAdapter (Middle)	11.62%	12.7M
ConvAdapter (Up)	11.78%	6.4M
ConvPass	15.13%	946.2K
BitFit	15.04%	192K
SSF	15.55%	73.7K
AdapterBias	14.81%	159.8K

Table 7.1: Performances of the different PETL alternatives on children ASR.

with 12.6 million parameters. The Scaled Adapter method exhibits a slightly reduced WER performance at 11.74% while maintaining the same parameter count of 12.6 million.

Turning attention to convolution-based Adapters, the replacement of either linear layer (W_{down} and W_{up}) results in decreased performance, yielding WERs of 12.02% and 11.78%, respectively, using both 6.4 million parameters. Interestingly, introducing a convolutional layer between these two linear layers proves to be the best-performing convolutional system, approaching the scores of the regular Adapter with 11.62% WER and using a slightly increased amount of parameters of 12.7 million. It is noteworthy that all these Conv-Adapter setups and the Scaled Adapters approach continue to outperform the fine-tuning of the entire model and are therefore valuable approaches for children's ASR PETL. On the other hand, the ConvPass method, where all linear layers of the Adapter were replaced by convolution layers, did not surpass the full fine-tuning, yielding a WER of 15.13% with 946.2 thousand parameters.

Moving to bias shift methods, such as BitFit, SSF, and AdapterBias, these approaches underperformed compared to the entire model fine-tuning, with respective WERs of 15.04%, 15.55%, and 14.81%. However, it is important to note that these methods use significantly fewer parameters, with 192, 73.7, and 159.8 thousand parameters, respectively.

In summary, the traditional residual Adapter remains the most effective PETL approach for children's ASR. This finding aligns with recent research, as highlighted in studies such as [283] and [248], which also affirm the superior performance of Adapters in PETL for ASR tasks. Moreover, our findings highlight a noticeable trade-off between the number of parameters and the WER. Specifically, approaches employing fewer than a million parameters did not exhibit comparable performances to entire model fine-tuning or other PETL methods with more parameters, as illustrated explicitly in Figure 7.3. This trade-off emphasises the need for more research on the development of PETL methods that can effectively use fewer parameters while simultaneously maintaining or enhancing performance in the domain of children's ASR.

7.3 Enhancing Parameter Efficiency through Transformer-based model Redundancy

7.3.1 Shared-Adapter

The primary objective of PETL is to either maintain or surpass the performance achieved through full fine-tuning of a pre-trained model while minimising the number of parameters employed during training. In the previous section, the efficiency of residual Adapters has been underscored. Remarkably, using only 10% of the total number of parameters compared to the fine-tuning of the entire model, these residual Adapters exhibit superior performances in the context of children's ASR. In addition, our findings, particularly from Chapter 4, highlighted the drawback of overparameterisation present in Transformer-based models.

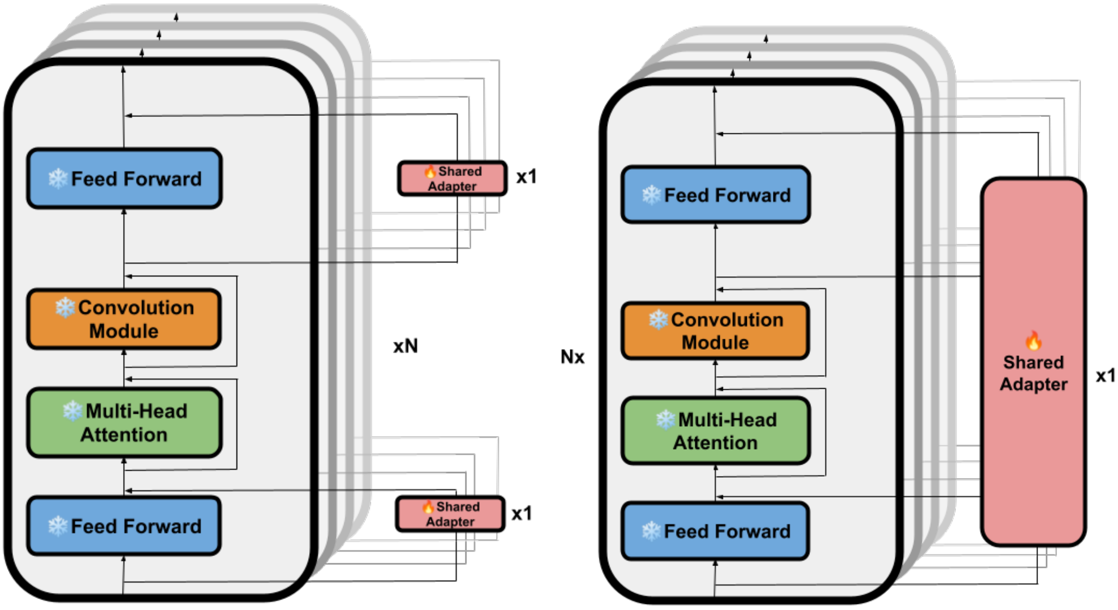
Additionally, the study by [291] found that the FFN modules, despite constituting a significant proportion of the model's parameters, were identified as highly redundant. This affirmation was corroborated by the work of [242], which establishes a connection between the FFN and attention mechanisms. The latter proposes that the FFN can be conceptualised as learnable key-value memories, with the weights of the first linear layer representing keys and those of the second layer corresponding to values. These keys are proficient at capturing salient patterns at each layer. Interestingly, they observed that the classes of patterns tend to overlap between neighbouring layers, indicating redundancy in the representations. This observation underscores the potential for optimising PETL methods by addressing and mitigating redundancy within FFN modules.

Building upon these observations, [291] modified the Transformer architecture by sharing and dropping the FFN across different layers. Their investigation confirms the substantial degree of redundancy between the FFNs of the Encoder and Decoder components. Consequently, they successfully eliminate the Decoder FFN while sharing a single FFN across the Encoder, achieving a noteworthy reduction in model parameters without significant compromise to accuracy.

Formally, with N_{enc} denoting the number of Encoder layers, the sharing of the FFN modules in the Encoder can be expressed as follows:

$$\mathrm{FFN}_i^{enc}(.) \overset{\text{tied}}{=} \mathrm{FFN}_{all}^{enc}(.), \forall i : 1 \leq i \leq N_{enc} \tag{7.7}$$

In light of the success observed with the shared FFN in the Encoder, we hypothesise that the presence of redundancy within the FFN might lead to a similar redundancy issue when employing one Adapter per FFN module. In other words, employing separate Adapters plugged into redundant FFN modules for different layers might also exhibit redundancy within the different Adapters. To address this concern, we introduced the *Shared Adapter* approach, wherein a single residual Adapter is used across all layers as



(a) Shared-Adapter setup in a Conformer model

(b) Extremly Shared-Adapter setup in a Conformer model

Figure 7.4: Overview of the Shared Adapters configurations

shown in figure 7.5. This approach aims to use the redundancy present in FFN layers modules to reduce the total amount of parameters used in Adapter-transfer. The formal expression of the Shared Adapter is presented as follows:

$$\text{Adapter}_i(.) \overset{\text{tied}}{=} \text{Shared-Adapter}_{all}(.), \forall i : 1 \leq i \leq N_{enc} \tag{7.8}$$

In order to evaluate our proposed approach, our initial focus is on comparing the Shared-Adapter configuration with full fine-tuning and the Traditional Adapter. Subsequently, we vary the hidden dimension of the Shared-Adapter to assess how the number of parameters and the hidden dimension influence the performances. This comprehensive analysis aims to provide insights into the effectiveness of Shared-Adapters across various scenarios and configurations. Finally, we assess the robustness of both Traditional and Shared-Adapters by examining their performances under different amounts of training data, simulating low and extremely low resource scenarios.

7.3.2 Experimental setup

In our analysis, we focus on evaluating the performance of the Shared-Adapter using the same setting as the TPA configuration for traditional Adapters. Especially because the TPA configuration was found to be the most effective in Conformer models. Specifically, we assess the Shared-Adapter within a Conformer

ASR model, where two Shared-Adapters are directly integrated into the two FFN modules for all layers as shown in Figure 7.4(a). The hidden dimension of the Shared-Adapters is set to 512. In terms of ASR, we used the same Conformer model as the rest of the thesis, which comprises 12 Conformer layers followed by 6 Transformer layers. As we solely want to asses the performances of Shared-Adapters to model acoustics variabilities of children's speech, we exclusively evaluate the Shared-Adapter in the Encoder only.

To quantify the reduction of the number of parameters used in the Shared-Adapter compared to the traditional Adapter, the following formula is used:

$$\text{Number of Parameters in Shared-Adapter} = \frac{\text{Number of Parameters of all traditional Adapters}}{\text{Number of Layers}} \tag{7.9}$$

In addition, we propose an extension of the Shared-Adapters concept within the TPA configuration. This novel approach, called *Extreme Shared-Adapter* involves using a single Adapter for all FFN modules instead of two. This configuration has the potential to further reduce the number of parameters by half compared to the use of two Shared-Adapters in the TPA setup. This Extreme Shared-Adapter configuration is represented in Figure 7.4(b). The *Extreme Shared-Adapter* exploration aims to test the limits of the parameter efficiency and accuracy balance of our approach within the context of children's ASR. For training all our different systems, we use 30 epochs with a learning rate set to 8×10^{-4}.

In terms of the number of parameters, while the traditional Adapter uses approximately 12.6 million parameters, the number of parameters employed by the Shared-Adapter transfer is approximately 1.1 million and the Extreme Shared-Adapter uses only 527.9 thousand parameters. This represents a substantial reduction, as the Shared-Adapter and Extreme Shared-Adapter configurations use only 8% and 3%, respectively, of the number of parameters of the traditional Adapters setup. Moreover, considering that the traditional Adapter already represents approximately 10% of the total parameters used in the entire model fine-tuning, the Shared-Adapter and Extreme Shared-Adapter configurations use only 1% and 0.5%, respectively, of the trainable parameters when compared to the entire model fine-tuning. This reduction in the number of parameters highlights the high parameter efficiency gains achieved by these novel configurations.

7.3.3 Results

7.3.3.A Shared-Adapter compared to other PETL methodologies

In our comprehensive analysis, we systematically evaluated the efficacy of our proposed Shared-Adapter configuration against established PETL methods mentioned in Section 7.2. The comparison, illustrated in Figure 7.5, highlights the remarkable performance of the Shared-Adapter, particularly in achieving a notable balance between parameter efficiency and accuracy. Notably, the Shared-Adapter outperforms full-model fine-tuning while using only 1% of the entire model's number of parameters, a significant

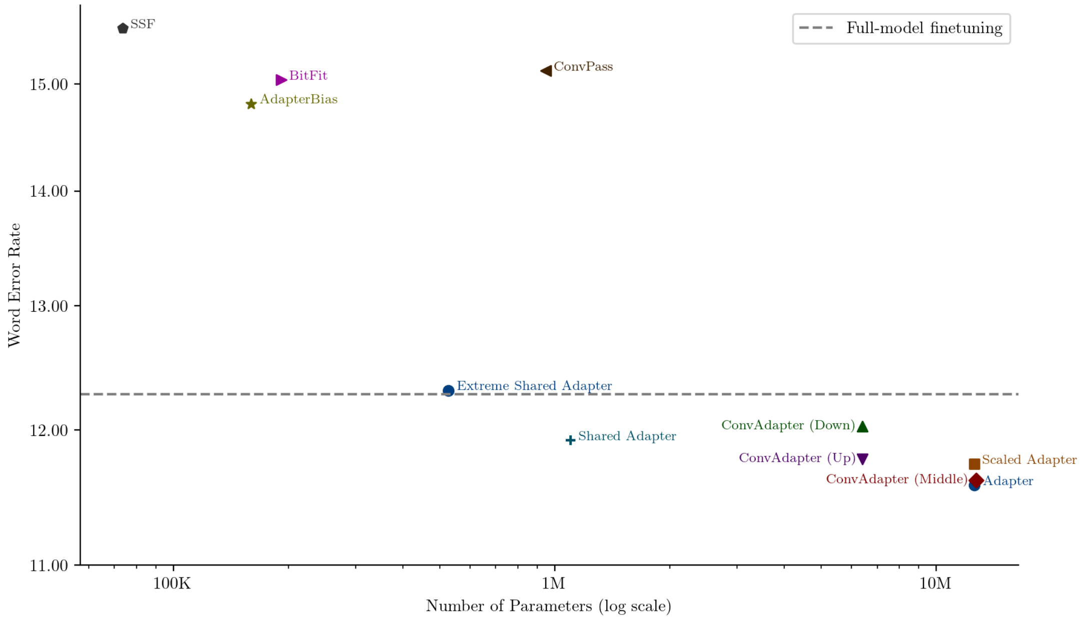


Figure 7.5: Different parameter efficient procedures for children ASR in Conformer model compared to the Shared-Adapters and Extremly Shared-Adapters

reduction compared to the 10% used by the Adapters. Although the Extreme Shared-Adapter exhibits a marginal degradation, surpassing the WER of the entire model fine-tuning, it still outperforms alternative methods such as ConvPass, BitFit, AdapterBias, and SSF by a considerable margin. These findings emphasise the efficiency of the Shared-Adapter as a novel parameter-efficient strategy, demonstrating a superior balance compared to existing methodologies. The precise results of the Shared-Adapter and Extreme Shared-Adapter will be presented in Table 7.2.

7.3.3.B Evaluating the hidden dimension and parameter influence on Shared-Adapter performances

The results in Table 7.2 present the performances of different configurations of the Shared-Adapter model with varying hidden dimensions compared to Traditional Adpater and full model fine-tuning. As mentioned in the previous section, Shared-Adapter and Extreme Shared-Adapter highlighted remarkable performances compared to Traditional Adapters and the full model fine-tuning while using fewer parameters. Additionally, as we decrease the hidden dimension to 128, the WER increases slightly to 12.74% WER, with a reduction in the number of trained parameters to 265.5K. Interestingly, as the hidden dimension increases from 128 to 4096, the WER consistently decreases. In particular, the Shared-Adapter with a hidden dimension of 4096 exhibits the best performance, achieving a WER of 11.83%, with a larger

Configuration	**WER ↓**	**Trained Params**
Full Fine-tuning	12.28%	109.1M
Traditional Adapter 512	**11.58%**	12.6M
Shared-Adapter 128	12.74%	265.5K
Shared-Adapter 256	12.34%	527.9K
Shared-Adapter 512	11.92%	1.1M
Shared-Adapter 1024	11.90%	2.1M
Shared-Adapter 2048	11.86%	4.2M
Shared-Adapter 4096	**11.83%**	8.4M
Shared-Adapter 6144	11.88%	12.6M
Extreme Shared-Adapter 512	12.31%	526.3K

Table 7.2: WER and Parameters for different Shared-Adapter hidden dimension

parameter count of 8.4M. However, as we push the hidden dimension further to 6144, giving an equivalent number of parameters as the Traditional Adapters, the performance starts to degrade. We also observed, a drop in score compared to the Traditional Adapters setup for the same number of parameters. This decline and drop in performance when using a large hidden dimension suggests a tipping point where the model's capacity to effectively learn relevant information becomes more challenging.

7.3.3.C Low resource and extremely low resource scenarios robustness

Number of hours	Fine-tune (109.1M)	TPA Adapter (12.6M)	Shared-Adapter (1.1M)
1h	16.24%	16.51%	17.20%
10h	14.00%	14.05%	14.28%
20h	13.43%	13.30%	13.52%
50h	12.94%	12.40%	12.57%
All (Ĩ13h)	12.28%	11.58%	11.92%

Table 7.3: WER for different training durations for the full model fine-tune, TPA Adapter, and Shared-Adapter

To assess the effectiveness of Adapter-transfer and the Shared-Adapter in settings with limited and extremely limited resources, we conducted training with varying amounts of training data. The results are outlined in Table 7.3. We considered three methodologies: Full model fine-tuning, Traditional Adapters using the TPA configuration, and Shared-Adapter, which also employed a TPA configuration with one Shared-Adapter per FFN module. The experiment covered scenarios with 50 and 20 hours, representing low-resource settings, and extremely low-resource scenarios with only 10 and 1 hour for training.

For full model fine-tuning, the WER ranges from 16.24% with 1 hour of training to 12.28% with the entire training set (approximately 113 hours). In parallel, Traditional Adapters show a WER reduction from 16.51% to 11.58% across the same durations. Notably, in extremely low-resource scenarios, Adapters and full transfer learning yield comparable results. Conversely, Adapters start to outperform the full model fine-tuning when the training duration exceeds 20 hours, reaffirming the efficiency of Adapter transfer for children's ASR.

The Shared-Adapter model exhibits competitive performances in all scenarios, with a WER decrease from 17.20% to a notable 11.92% as training duration extends. This establishes Shared-Adapter as a robust PETL approach even in low-resource and extremely low-resource scenarios of children's speech.

7.4 Summary and discussion

In this chapter, we addressed to the following research question: *Can we further improve the parameter-efficiency with other architectures?*. The comprehensive evaluation of various parameter-efficient transfer learning alternatives from the existing literature revealed that these alternatives did not surpass the performance achieved by traditional Adapter modules. Furthermore, our results on children's ASR highlighted the presence of an important tradeoff between accuracy and parameter efficiency.

Crucially, the optimal balance was observed when PETL methods used around 10% of the total parameters of the entire pre-trained model. This parameter-efficient configuration consistently resulted in WER scores below those obtained through the full-model fine-tuning. However, reducing the proportion of parameters trained to less than 10% of the full model led to a notable deterioration in WER scores, underlining the tradeoff between parameter efficiency gains and preservation of accuracy in children's ASR systems.

Additionally, we introduced a novel PETL approach, the Shared-Adapter. Leveraging the inherent redundancy of FFN modules in between the different Transformer layers. Our experimental evaluations demonstrated that Shared-Adapters represent a breakthrough in parameter efficiency while maintaining high accuracy in children's ASR. Notably, while conventional PETL methods typically required around 10% of the total amount of the entire model parameters, Shared-Adapters excelled by achieving comparable accuracy with only 1% of the parameters. Pushing the boundaries even further, we explored the extreme scenario of using the Extreme Shared-Adapters configuration with a mere 0.5% of the model's total parameters, yet still achieving performance levels similar to full-model fine-tuning. Our approach eliminates the need for the aforementioned tradeoff, offering a pathway to obtain superior parameter efficiency without compromising accuracy.

The remarkable success of the Shared-Adapter approach underscores the significant redundancy of the FFN modules across different layers. This new understanding is pivotal for advancing the development of more efficient, computationally compact ASR models. Future research directions could explore the creation of a novel Transformer architecture designed to address the observed redundancy, providing an architecture that is inherently easier to fine-tune for children's ASR or speaker-specific tasks. Additionally, investigating a new approach that incorporates both Shared and non-shared Adapters at the same time represents another avenue for potential advancements in the field. These avenues hold promise for enhancing the efficiency and robustness of ASR models, particularly in contexts with limited data, such as children's speech recognition.

8

Conclusions

Contents

The primary objective of this thesis was to improve ASR for children's speech. Initially, the focus was on the necessity of developing a reliable foundational children's ASR model to support the development of SLT for children with pathological speech. However, the focus of the thesis shifted towards solely the improvement of healthy children's ASR. This redirection expanded the potential applications of our research, moving beyond the initial scope of possible SLT applications. These new potential applications include personal assistants, automatic reading tutors, and voice interactions with computer interfaces. However, we note that recognising the unique characteristics and challenges of children's ASR is crucial for developing a foundational ASR model that would be robust across different contexts, including pathological children's speech.

This thesis has made significant contributions to the field of children's ASR, with a particular emphasis on improving traditional knowledge transfer methods which represented the state-of-the-art methods for children's ASR. The proposed improvements in this thesis aim to make these ASR models more robust, granular, and parameter-efficient, both in hybrid and end-to-end frameworks. In this concluding chapter, we provide a comprehensive summary of the various works developed during this thesis, highlighting key discoveries and contributions to the field of children's ASR.

8.1 Summary of the work carried out during the thesis

In the first part of the thesis, we established the context of our research and outlined the challenges associated with children's speech recognition. The primary challenge arises from the high variability exhibited in children's speech, particularly the acoustics. Notable features of children's speech include frequency ranges that are shifted and broadened compared to adult speech, along with significant inter- and intra-speaker variability. Furthermore, the process of language acquisition for young children adds an additional layer of complexity, making it challenging for both human listeners and automated systems to accurately recognise children's speech. Addressing this complexity necessitates a substantial amount of data, constituting a second major challenge due to the scarcity of available children's speech data. In this initial section, we present a non-exhaustive list of children's speech corpora available in the literature, representing the most extensive compilation to the best of our knowledge. Additionally, we reviewed various approaches employed in the literature to tackle the diverse challenges inherent in children's ASR.

In the second part of the thesis, our focus was on the development and exploration of a hybrid HMM-DNN ASR system, specifically designed for low-resource children's ASR, with a particular emphasis on the European Portuguese language. Our efforts centered around the evaluation of various knowledge transfer approaches, to asses their efficiency for children's ASR. Among the different approaches considered, transfer learning emerged as the most effective for systems dedicated solely to recognising children's speech. Additionally, multi-task learning proved effective when the system needed to recognise both children and adult speech simultaneously. We also introduced a novel approach, the *multilingual transfer* *learning*, which involves both multi-task and transfer learning. Our findings demonstrate that training a multilingual children's ASR system serves as a superior initialisation for subsequent transfer learning on a target children's dataset, especially in low-resourced scenarios.

In the subsequent part of our research, we transitioned towards the end-to-end paradigm, seeking to enhance the current state-of-the-art approaches. Instead of transfer learning over the entire model, we proposed a more granular evaluation. Our investigation revealed that the Encoder plays a pivotal role in the fine-tuning process for end-to-end children's ASR. This aligns with the understanding that in the context of children's speech, acoustic variability highly contributes to the degradation of recognition accuracy, more than linguistic factors. Furthermore, our findings highlighted that higher layers, closer to the output of the Encoder, yield more substantial benefits in fine-tuning compared to lower layers, near the input of the Encoder. These insights offer valuable recommendations for the development of children's ASR models through transfer learning. In this section of the thesis, we also introduced the *partial fine-tuning* approach for Transformer-based architectures. We identified that fine-tuning only specific components of the network outperformed entire model fine-tuning, with the FFN component being recognised as the most crucial module with a relative WER improvement around 9% for both Transformer and Conformer architectures.

Next, motivated by the need for parameter efficiency knowledge transfer, especially in scenarios with limited training data, we delved into the exploration of the use of Adapter modules. Adapter modules, comprising two linear layers integrated into a pre-trained frozen model, serve as a mechanism for knowledge transfer while preserving the weights and knowledge contained in the pre-trained model. We evaluated various configurations within both Transformer and Conformer architectures. Among all configurations tested, the parallel configuration, and its Conformer extension known as TPA, emerged as the most effective for transferring knowledge to children's speech. Remarkably, these configurations outperformed entire model fine-tuning by achieving better results while using only 10% of the number of parameters involved in entire model fine-tuning. This evaluation demonstrated the viability of Adapters in the context of children's ASR, suggesting their applicability for more precise adaptation. For this purpose, we introduced an unsupervised procedure in which utterances were clustered using k-means applied to their respective x-vectors. The rationale behind this approach lies in the notion that utterances sharing acoustic similarities, as captured by the x-vector extractor, would benefit more from an Adapter specifically trained on similar utterances rather than a general children's Adapter. Our experiments, involving the manipulation of cluster numbers, yielded additional improvements in children's speech recognition results.

Expanding on the efficiency of Adapters in bridging the gap between the source and target domains for children's speech, we employed Adpaters to improve data augmentation with imperfect data for children's ASR. Specifically, we introduced the *DWAT*, incorporating TTS data as synthetic speech. Given that synthetic speech often exhibits acoustics mismatch with real speech, there was a crucial need to reduce this gap, which our DWAT approach helped to address. This two-step procedure involved the initial training of Adapter modules using imperfect TTS data, followed by the fine-tuning of Adapters and the entire model weights using both synthetic and real data. The data underwent a distinctive dual-pathway, with synthetic speech passing through the Adapters while real speech skipped this them. The DWAT approach resulted in notable improvements over baseline and previous techniques across both Transformer and Conformer architectures, showcasing the efficacy of our approach. Additionally, we extended speaker embedding filtering of imperfect synthetic data by incorporating x-vectors instead of i-vectors, using cosine similarity between the reference and generated utterances to discard weak similarities that may represent incorrectly generated utterances.

Finally, motivated by the different successes of Adapters, both in Adapter-transfer and the DWAT approach, we evaluated different alternatives present in the literature. Our observations for children's ASR revealed that, among the various parameter-efficient methods, traditional Adapters remained the most effective. We noted a tradeoff between accuracy and parameter efficiency, some methods were highly parameter-efficient but led to significantly degraded results, while others were less parameter-efficient but yielded comparable or better results than the entire model fine-tuning. To address this tradeoff, we proposed a novel approach that leverages the high redundancy present in the FFN components of Transformer-based models. Our *Shared-Adapters*, where only one Adapter is shared across all layers instead of having one per layer, outperformed the entire model fine-tuning. While facing a minimal score degradation compared to the traditional Adapter, our Shared-Adapter was trained on a significantly smaller amount of parameters than any previous approach with this range of score. Using less than 1% of the entire model's number of parameters compared to the traditional Adapter's 10%, the Shared-Adapters provide a promising solution to the tradeoff challenge and emerge as an excellent candidate for achieving better parameter efficiency transfer in children's ASR.

8.2 Perspectives

At the conclusion of this thesis, several promising avenues for further improvement of children's ASR have been identified. These perspectives could provide a foundation for future research endeavours in the field of children's ASR, aiming to advance our understanding and enhance the capabilities of ASR systems tailored for young speakers. These perspectives include:

Extension of the Double-way Adapter Transfer to other domains The DWAT approach has proven to be effective in enhancing the overall fine-tuning process by using Adapters to bridge the gap between two distinct domains, particularly between synthetic and real data. An interesting avenue for further research would be to extend this approach to novel domains, such as adult and children's speech, exploring the adaptability and efficacy of Adapters in diverse contexts. Furthermore, in our work, we primarily employed entire model fine-tuning; however, our findings of chapter 4 suggest that a more granular fine-tuning approach could yield even better results. This indicates the importance of investigating the fine-tuning process of the DWAT approach at a more detailed level, focusing on specific components or layers rather than the entire network.

Exploration of novel non-redundant Transformer-based architecture for efficient fine-tuning The success of our Shared-Adapter shed light on the redundancy present in the FFN components of the different layers of Transformer-based models. As we showed during this thesis, fine-tuning overparameterised networks with a small amount of data could lead to overfitting or decreased WER scores. Therefore, it would be interesting to develop a novel architecture of Transformer-based ASR models that would not be as redundant. Either by using a shared FFN components in a similar way as [291]. This model would be easier to fine-tune for low-resource scenarios. In addition, mixed Shared and layers-wise FFN modules could be as well an interesting avenue of research.

Furhter exploration of Shared Adapters The good results observed with Shared-Adapters paves the way for further developments, suggesting potential exploration of hybrid approaches that combine shared Adapters with layer-wise Adapters. This mixed approach could offer a more nuanced strategy for parameter-efficient transfer learning in children's ASR. Exploring the synergies between different Adapter configurations could contribute to refining the adaptability of models to children's speech while maintaining high efficiency.

Developement of Speech and Language Technology based on improved children's Automatic Speech Recognition Collecting pathological speech data for children would be a valuable and important initiative. The availability of such data will contribute to the development of technology that is specifically tailored for children with speech disorders using our improved children's ASR systems as a foundation. This can lead to improved speech and language technology applications, allowing for more inclusive and effective solutions for children with pathological speech.

Bibliography

[1] M. Gerosa, D. Giuliani, S. Narayanan, and A. Potamianos, "A review of asr technologies for children's speech," in *Proceedings of the 2nd Workshop on Child, Computer and Interaction*, ser. WOCCI '09. New York, NY, USA: Association for Computing Machinery, 2009. [Online]. Available: https://doi.org/10.1145/1640377.1640384

[2] S. Lee, A. Potamianos, and S. 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. [Online]. Available: https://doi.org/10.1121/1.426686

[3] D. H. Klatt, "Review of the arpa speech understanding project," *The Journal of the Acoustical Society of America*, vol. 62, no. 6, pp. 1345–1366, 1977.

[4] E. Kiktova, M. Lojka, M. Pleva, J. Juhar, and A. Cizmar, "Comparison of different feature types for acoustic event detection system," in *Multimedia Communications, Services and Security: 6th International Conference, MCSS 2013, Krakow, Poland, June 6-7, 2013. Proceedings 6*. Springer, 2013, pp. 288–297.

[5] R. Weide *et al.*, "The carnegie mellon pronouncing dictionary," *release 0.6, www. cs. cmu. edu*, 1998.

[6] P. G. Shivakumar and P. Georgiou, "Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations," 2018.

[7] F. Wu, L. P. García-Perera, D. Povey, and S. Khudanpur, "Advances in automatic speech recognition for child speech using factored time delay neural network." in *Interspeech*, 2019, pp. 1–5.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[9] W. J. Levelt, *Speaking: From intention to articulation*. MIT press, 1993.

[10] L. Black, A. Vahratian, and H. Hoffman, "Communication disorders and use of intervention services among children aged 3–17 years: United states, 2012; us department of health and human services, centers for disease control and prevention," *National Center for Health Statistics: Atlanta, GA, USA*, 2015.

[11] D. Langbecker, C. L. Snoswell, A. C. Smith, J. Verboom, and L. J. Caffery, "Long-term effects of childhood speech and language disorders: A scoping review," *South African Journal of Childhood Education*, vol. 10, no. 1, pp. 1–13, 2020.

[12] D. Hilty, S. Chan, J. Torous, J. Mahautmr, and D. Mucic, "New frontiers in healthcare and technology: Internet-and web-based mental options emerge to complement in-person and telepsychiatric care options," *J Health Med Informatics*, vol. 6, no. 4, pp. 1–14, 2015.

[13] J. E. Barnett, "Utilizing technological innovations to enhance psychotherapy supervision, training, and outcomes." *Psychotherapy*, vol. 48, no. 2, p. 103, 2011.

[14] M. C. Hughes, J. M. Gorman, Y. Ren, S. Khalid, and C. Clayton, "Increasing access to rural mental health care using hybrid care that includes telepsychiatry." *Journal of Rural Mental Health*, vol. 43, no. 1, p. 30, 2019.

[15] V. Mendoza Ramos, "The added value of speech technology in clinical care of patients with dysarthria," Ph.D. dissertation, University of Antwerp, 2022.

[16] R. Brewer, L. Anthony, Q. Brown, G. Irwin, J. Nias, and B. Tate, "Using gamification to motivate children to complete empirical studies in lab environments," in *Proceedings of the 12th international conference on interaction design and children*, 2013, pp. 388–391.

[17] A. Potamianos and S. Narayanan, "Robust recognition of children's speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 603–616, 2003.

[18] Y. Li, Z. Zhao, O. Klejch, P. Bell, and C. Lai, "Asr and emotional speech: A word-level investigation of the mutual impact of speech and emotion recognition," *arXiv preprint arXiv:2305.16065*, 2023.

[19] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[20] B. King, I.-F. Chen, Y. Vaizman, Y. Liu, R. Maas, S. H. K. Parthasarathi, and B. Hoffmeister, "Robust speech recognition via anchor word representations," 2017.

[21] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *The Journal of the Acoustical Society of America*, vol. 24, no. 2, pp. 175–184, 1952. [Online]. Available: https://doi.org/10.1121/1.1906875

[22] Q. Li and M. J. Russell, "Why is automatic recognition of children's speech difficult?" in *Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, 2001, pp. 2671–2674.

[23] D. CRYSTAL, "A dictionary of linguistics and phonetics (5th edn.). oxford: Blackwell publishing, 2003. pp. 508. isbn 0 631 22664 8," *Journal of the International Phonetic Association*, vol. 34, pp. 100 – 101, 01 2004.

[24] L. C. Moats and S. Brady, *Speech to print: Language essentials for teachers.* Paul H. Brookes Pub., 2000.

[25] H. Tulsiani, P. Swarup, and P. Rao, "Acoustic and language modeling for children's read speech assessment," in *2017 Twenty-third National Conference on Communications (NCC).* IEEE, 2017, pp. 1–6.

[26] H. H. Clark and E. V. Clark, "Psychology and language," 1977.

[27] A. Potamianos and S. Narayanan, "Spoken dialog systems for children," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 1, 06 1998, pp. 197 – 200 vol.1.

[28] S. Das, D. Nix, and M. Picheny, "Improvements in children's speech recognition performance," *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, vol. 1, pp. 433–436 vol.1, 1998.

[29] S. S. Gray, D. Willett, J. Lu, J. Pinto, P. Maergner, and N. Bodenstab, "Child automatic speech recognition for us english: child interaction with living-room-electronic-devices," in *WOCCI*, 2014.

[30] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning.* PMLR, 2023, pp. 28 492–28 518.

[31] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *ICASSP*, 2015, pp. 5206–5210.

[32] G. Chen, S. Chai, G. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang *et al.*, "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," *arXiv preprint arXiv:2106.06909*, 2021.

[33] W. Ward, R. Cole, D. Bolaños, C. Buchenroth-Martin, E. Svirsky, and T. B. Weston, "My science tutor: A conversational multimedia virtual tutor." *Journal of Educational Psychology*, vol. 105, pp. 1115–1125, 2013.

[34] N. F. Chen, R. Tong, D. Wee, P. X. Lee, B. Ma, and H. Li, "Singakids-mandarin: Speech corpus of singaporean children speaking mandarin chinese." in *Interspeech*, 2016, pp. 1545–1549.

[35] B. Ahmed, K. Ballard, D. Burnham, T. Sirojan, H. Mehmood, D. Estival, E. Baker, F. Cox, J. Arciuli, T. Benders *et al.*, "Auskidtalk: an auditory-visual corpus of 3-to 12-year-old australian children's speech," in *Annual Conference of the International Speech Communication Association (22nd: 2021)*. International Speech Communication Association, 2021, pp. 3680–3684.

[36] M. Eskenazi, J. Mostow, and D. Graff, "The cmu kids speech corpus," *Corpus of children's read speech digitized and transcribed on two CD-ROMs, with assistance from Multicom Research and David Graff. Published by the Linguistic Data Consortium, University of Pennsylvania*, 1997.

[37] K. Shobaki, J.-P. Hosom, and R. Cole, "The ogi kids' speech corpus and recognizers," in *Proc. of ICSLP*, 2000, pp. 564–567.

[38] M. Russell, "The pf-star british english children's speech corpus," 2006.

[39] H. Liao, G. Pundak, O. Siohan, M. Carroll, N. Coccaro, Q.-M. Jiang, T. N. Sainath, A. Senior, F. Beaufays, and M. Bacchiani, "Large vocabulary automatic speech recognition for children," in *Interspeech*, 2015.

[40] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris *et al.*, "Automatic speech recognition and speech variability: A review," *Speech communication*, vol. 49, no. 10-11, pp. 763–786, 2007.

[41] S. Karpagavalli and E. Chandra, "A review on automatic speech recognition architecture and approaches," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 9, no. 4, pp. 393–404, 2016.

[42] S. J. Arora and R. P. Singh, "Automatic speech recognition: a review," *International Journal of Computer Applications*, vol. 60, no. 9, 2012.

[43] K. H. Davis, R. Biddulph, and S. Balashek, "Automatic recognition of spoken digits," *The Journal of the Acoustical Society of America*, vol. 24, no. 6, pp. 637–642, 1952. [Online]. Available: https://doi.org/10.1121/1.1906946

[44] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey *et al.*, "The htk book," *Cambridge university engineering department*, vol. 3, no. 175, p. 12, 2002.

[45] P. C. Woodland and S. J. Young, "The htk tied-state continuous speech recogniser." in *Eurospeech*, 1993.

[46] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[47] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[48] J. Baker, "The dragon system–an overview," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 1, pp. 24–29, 1975.

[49] A. L. Bizzocchi, "How many phonemes does the english language have?" *International Journal on Studies in English Language and Literature (IJSELL)*, vol. 5, no. 10, pp. 36–46, 2017.

[50] R. Schwartz, Y. Chow, O. Kimball, S. Roucos, M. Krasner, and J. Makhoul, "Context-dependent modeling for acoustic-phonetic recognition of continuous speech," in *ICASSP'85. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 10. IEEE, 1985, pp. 1205–1208.

[51] L. R. Bahl, P. V. deSouza, P. Gopalakrishnan, D. Nahamoo, and M. Picheny, "Context dependent modeling of phones in continuous speech using decision trees," in *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991*, 1991.

[52] H. A. Bourlard and N. Morgan, *Connectionist speech recognition: a hybrid approach.* Springer Science & Business Media, 2012, vol. 247.

[53] H. Meinedo, D. Caseiro, J. Neto, and I. Trancoso, "Audimus. media: a broadcast news speech recognition system for the european portuguese language," in *International Workshop on Computational Processing of the Portuguese Language.* Springer, 2003, pp. 9–17.

[54] K. J. Lang, A. H. Waibel, and G. E. Hinton, "A time-delay neural network architecture for isolated word recognition," *Neural networks*, vol. 3, no. 1, pp. 23–43, 1990.

[55] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," 2014.

[56] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," in *Backpropagation.* Psychology Press, 2013, pp. 35–61.

[57] K. Yao and G. Zweig, "Sequence-to-sequence neural net models for grapheme-to-phoneme conversion," *arXiv preprint arXiv:1506.00196*, 2015.

[58] G. Sidorov, F. Velasquez, E. Stamatatos, A. Gelbukh, and L. Chanona-Hernández, "Syntactic n-grams as machine learning features for natural language processing," *Expert Systems with Applications*, vol. 41, no. 3, pp. 853–860, 2014.

[59] S. Vishnoi, P. Garg, and P. Arora, "Physicochemical n-grams tool: A tool for protein physicochemical descriptor generation via chou's 5-step rule," *Chemical Biology & Drug Design*, vol. 95, no. 1, pp. 79–86, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/cbdd.13617

[60] V. H. Nguyen, H. T. Nguyen, H. N. Duong, and V. Snasel, "n-Gram-Based Text Compression," *Computational Intelligence and Neuroscience*, vol. 2016, p. 9483646, Nov. 2016, publisher: Hindawi Publishing Corporation. [Online]. Available: https://doi.org/10.1155/2016/9483646

[61] S. Chen and R. Rosenfeld, "A survey of smoothing techniques for me models," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 1, pp. 37–50, 2000.

[62] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[63] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[64] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.

[65] V. Valtchev, J. Odell, P. Woodland, and S. Young, "A novel decoder design for large vocabulary recognition," in *Proceedings of ICSLP*, 1994.

[66] X. Aubert and H. Ney, "Large vocabulary continuous speech recognition using word graphs," in *1995 International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 1995, pp. 49–52.

[67] M. Mohri, "Finite-state transducers in language and speech processing," *Computational linguistics*, vol. 23, no. 2, pp. 269–311, 1997.

[68] D. Caseiro and I. Trancoso, "Using dynamic wfst composition for recognizing broadcast news," in *Seventh International Conference on Spoken Language Processing*, 2002.

[69] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on*

automatic speech recognition and understanding, no. CONF. IEEE Signal Processing Society, 2011.

[70] F. Richardson, M. Ostendorf, and J. R. Rohlicek, “Lattice-based search strategies for large vocabulary speech recognition,” in *1995 International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 1995, pp. 576–579.

[71] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, “Deep speech: Scaling up end-to-end speech recognition,” *arXiv preprint arXiv:1412.5567*, 2014.

[72] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang, S. Watanabe, T. Yoshimura, and W. Zhang, “A comparative study on transformer vs rnn in speech applications,” in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 449–456.

[73] D. Wang, X. Wang, and S. Lv, “An overview of end-to-end automatic speech recognition,” *Symmetry*, vol. 11, p. 1018, 08 2019.

[74] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML ’06. New York, NY, USA: Association for Computing Machinery, 2006, p. 369–376. [Online]. Available: https://doi.org/10.1145/1143844.1143891

[75] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” *Advances in neural information processing systems*, vol. 27, 2014.

[76] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *arXiv preprint arXiv:1409.0473*, 2014.

[77] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3156–3164.

[78] O. Vinyals and Q. Le, “A neural conversational model,” *arXiv preprint arXiv:1506.05869*, 2015.

[79] R. Nallapati, B. Zhou, C. Gulcehre, B. Xiang *et al.*, “Abstractive text summarization using sequence-to-sequence rnns and beyond,” *arXiv preprint arXiv:1602.06023*, 2016.

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

[81] Z. Tüske, G. Saon, and B. Kingsbury, "On the limit of english conversational speech recognition," *arXiv preprint arXiv:2105.00982*, 2021.

[82] D. Bermuth, A. Poeppel, and W. Reif, "Scribosermo: fast speech-to-text models for german and other languages," *arXiv preprint arXiv:2110.07982*, 2021.

[83] W. Chan, D. Park, C. Lee, Y. Zhang, Q. Le, and M. Norouzi, "Speechstew: Simply mix all available speech recognition data to train one large neural network," *arXiv preprint arXiv:2104.02133*, 2021.

[84] S. Ghai and R. Sinha, "Exploring the role of spectral smoothing in context of children's speech recognition," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[85] S. Ghai, "Addressing pitch mismatch for children's automatic speech recognition," Ph.D. dissertation, 2011.

[86] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech." *The Journal of the Acoustical Society of America*, vol. 87 4, pp. 1738–52, 1990.

[87] S. Shahnawazuddin, Ankita, A. Kumar, and H. K. Kathania, "Gammatone-filterbank based pitch-normalized cepstral coefficients for zero-resource children's asr," in *International Conference on Speech and Computer*. Springer, 2023, pp. 494–505.

[88] S. P. Dubagunta, S. Hande Kabil, and M. Magimai.-Doss, "Improving children speech recognition through feature learning from raw speech signal," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5736–5740.

[89] J. Fainberg, O. Klejch, E. Loweimi, P. Bell, and S. Renals, "Acoustic model adaptation from raw waveforms with sincnet," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 897–904.

[90] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," 2019.

[91] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 1. IEEE, 1996, pp. 353–356.

[92] R. Serizel and D. Giuliani, "Vocal tract length normalisation approaches to dnn-based children's and adults' speech recognition," in *SLT Workshop*, 2014, pp. 135–140.

[93] F. Claus, H. Gamboa Rosales, R. Petrick, H.-U. Hain, and R. Hoffmann, "A survey about asr for children," in *Speech and Language Technology in Education*, 2013.

[94] A. Potamianos, S. Narayanan, and S. Lee, "Automatic speech recognition for children," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[95] A. Potamianos and R. C. Rose, "On combining frequency warping and spectral shaping in hmm based speech recognition," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE, 1997, pp. 1275–1278.

[96] G. Yeung and A. Alwan, "A frequency normalization technique for kindergarten speech recognition inspired by the role of f0 in vowel perception," *Interspeech 2019*, 2019.

[97] S. Shahnawazuddin, R. Sinha, and G. Pradhan, "Pitch-normalized acoustic features for robust children's speech recognition," *IEEE Signal Processing Letters*, vol. 24, no. 8, pp. 1128–1132, 2017.

[98] S. Shahnawazuddin, A. Dey, and R. Sinha, "Pitch-adaptive front-end features for robust children's asr," in *INTERSPEECH*, 2016.

[99] H. Kathania, S. Kadiri, P. Alku, and M. Kurimo, "A formant modification method for improved asr of children's speech," *Speech Communication*, vol. 136, pp. 98–106, 01 2022.

[100] U. L. Kumar, M. Kurimo, and H. K. Kathania, "Effect of linear prediction order to modify formant locations for children speech recognition," in *International Conference on Speech and Computer*. Springer, 2023, pp. 483–493.

[101] H. K. Kathania, S. Shahnawazuddin, W. Ahmad, N. Adiga, S. K. Jana, and A. B. Samaddar, "Improving children's speech recognition through time scale modification based speaking rate adaptation," in *2018 International Conference on Signal Processing and Communications (SPCOM)*. IEEE, 2018, pp. 257–261.

[102] R. Duan and N. F. Chen, "Senone-aware adversarial multi-task training for unsupervised child to adult speech adaptation," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7758–7762.

[103] L. Rumberg, H. Ehlert, U. Lüdtke, and J. Ostermann, "Age-invariant training for end-to-end child speech recognition using adversarial multi-task learning," *Proc. Interspeech 2021*, pp. 3850–3854, 2021.

[104] A. Senior and I. Lopez-Moreno, "Improving dnn speaker independence with i-vector inputs," in *ICASSP*, 2014, pp. 225–229.

[105] 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.

[106] H. K. Kathania, S. Shahnawazuddin, N. Adiga, and W. Ahmad, "Role of prosodic features on children's speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5519–5523.

[107] V. Kadyan, T. Hasija, and A. Singh, "Prosody features based low resource punjabi children asr and t-nt classifier using data augmentation," *Multimedia Tools and Applications*, vol. 82, no. 3, pp. 3973–3994, 2023.

[108] J. Wilpon and C. Jacobsen, "A study of speech recognition for children and the elderly," in *ICASSP*, vol. 1, 1996, pp. 349–352 vol. 1.

[109] R. Gale, L. Chen, J. Dolata, J. Van Santen, and M. Asgari, "Improving asr systems for children with autism and language impairment using domain-focused dnn transfer techniques," in *Interspeech*, vol. 2019. NIH Public Access, 2019, p. 11.

[110] A. Hagen, B. Pellom, and R. Cole, "Highly accurate children's speech recognition for interactive reading tutors using subword units," *Speech Communication*, vol. 49, no. 12, pp. 861–873, 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639307000878

[111] P. G. Shivakumar, A. Potamianos, S. Lee, and S. S. Narayanan, "Improving speech recognition for children using acoustic adaptation and pronunciation modeling." in *WOCCI*, 2014, pp. 15–19.

[112] Q. Li and M. J. Russell, "An analysis of the causes of increased error rates in children's speech recognition," in *Seventh International Conference on Spoken Language Processing*, 2002.

[113] V. Bhardwaj, M. T. Ben Othman, V. Kukreja, Y. Belkhier, M. Bajaj, B. S. Goud, A. U. Rehman, M. Shafiq, and H. Hamam, "Automatic speech recognition (asr) systems for children: A systematic literature review," *Applied Sciences*, vol. 12, no. 9, p. 4419, 2022.

[114] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth annual conference of the international speech communication association*, 2015.

[115] M. Kumar, S. H. Kim, C. Lord, T. D. Lyon, and S. Narayanan, "Leveraging linguistic context in dyadic interactions to improve automatic speech recognition for children," *Computer speech & language*, vol. 63, p. 101101, 2020.

[116] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks." in *Interspeech*, 2018, pp. 3743–3747.

[117] 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.

[118] P. Gurunath Shivakumar and S. Narayanan, "End-to-end neural systems for automatic children speech recognition: An empirical study," *Computer Speech & Language*, vol. 72, p. 101289, 2022.

[119] G. Chen, X. Na, Y. Wang, Z. Yan, J. Zhang, S. Ma, and Y. Wang, "Data augmentation for children's speech recognition–the" ethiopian" system for the slt 2021 children speech recognition challenge," *arXiv preprint arXiv:2011.04547*, 2020.

[120] S.-I. Ng, W. Liu, Z. Peng, S. Feng, H.-P. Huang, O. Scharenborg, and T. Lee, "The cuhk-tudelft system for the slt 2021 children speech recognition challenge," *arXiv preprint arXiv:2011.06239*, 2020.

[121] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell," *arXiv preprint arXiv:1508.01211*, 2015.

[122] S. Targ, D. Almeida, and K. Lyman, "Resnet in resnet: Generalizing residual architectures," *arXiv preprint arXiv:1603.08029*, 2016.

[123] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[124] M. Qian, I. McLoughlin, W. Quo, and L. Dai, "Mismatched training data enhancement for automatic recognition of children's speech using dnn-hmm," in *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2016, pp. 1–5.

[125] J. Fainberg, P. Bell, M. Lincoln, and S. Renals, "Improving children's speech recognition through out-of-domain data augmentation." in *Interspeech*, 2016, pp. 1598–1602.

[126] M. Matassoni, R. Gretter, D. Falavigna, and D. Giuliani, "Non-native children speech recognition through transfer learning," in *ICASSP*, 2018, pp. 6229–6233.

[127] T. Nagano, T. Fukuda, M. Suzuki, and G. Kurata, "Data augmentation based on vowel stretch for improving children's speech recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 502–508.

[128] G. Yeung, R. Fan, and A. Alwan, "Fundamental frequency feature normalization and data augmentation for child speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6993–6997.

[129] M. Dua, V. Kadyan, N. Banthia, A. Bansal, and T. Agarwal, “Spectral warping and data augmentation for low resource language asr system under mismatched conditions,” *Applied Acoustics*, vol. 190, p. 108643, 2022.

[130] Z. Shuyang, M. Singh, A. Woubie, and R. Karhila, “Data augmentation for children asr and child-adult speaker classification using voice conversion methods.”

[131] P. Sheng, Z. Yang, and Y. Qian, “Gans for children: A generative data augmentation strategy for children speech recognition,” in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 129–135.

[132] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, “Natural tts synthesis by conditioning wavenet on mel spectrogram predictions,” in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[133] J. Kim, J. Kong, and J. Son, “Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.

[134] A. Laptev, R. Korostik, A. Svischev, A. Andrusenko, I. Medennikov, and S. Rybin, “You do not need more data: Improving end-to-end speech recognition by text-to-speech data augmentation,” in *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE, 2020, pp. 439–444.

[135] W. Wang, Z. Zhou, Y. Lu, H. Wang, C. Du, and Y. Qian, “Towards data selection on tts data for children’s speech recognition,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6888–6892.

[136] F.-H. Liu, Y. Gao, L. Gu, and M. Picheny, “Noise robustness in speech to speech translation,” in *Eighth European Conference on Speech Communication and Technology*, 2003.

[137] C. Ris and S. Dupont, “Assessing local noise level estimation methods: Application to noise robust asr,” *Speech Communication*, vol. 34, no. 1, pp. 141–158, 2001, noise Robust ASR. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639300000510

[138] L. Gelin, M. Daniel, T. Pellegrini, and J. Pinquier, “Babble noise augmentation for phone recognition applied to children reading aloud in a classroom environment,” in *Speech in Noise Workshop (SPiN)*, 2020.

[139] L. Couvreur and C. Couvreur, "On the use of artificial reverberation for asr in highly reverberant environments," in *Proc. 2nd IEEE Benelux Signal Processing Symposium (SPS-2000), Hilvarenbeek, The Netherlands.* Citeseer, 2000, pp. S001–S004.

[140] J. Malek, J. Zdansky, and P. Cerva, "Robust automatic recognition of speech with background music," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2017, pp. 5210–5214.

[141] T.-H. Lo, F.-A. Chao, S.-Y. Weng, and B. Chen, "The ntnu system at the interspeech 2020 non-native children's speech asr challenge," *arXiv preprint arXiv:2005.08433*, 2020.

[142] V. P. Singh, H. Sailor, S. Bhattacharya, and A. Pandey, "Spectral modification based data augmentation for improving end-to-end asr for children's speech," *arXiv preprint arXiv:2203.06600*, 2022.

[143] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, vol. 117, 2013, p. 21.

[144] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple augmentation method for automatic speech recognition," in *INTERSPEECH*, 2019.

[145] L. Gelin, T. Pellegrini, J. Pinquier, and M. Daniel, "Simulating reading mistakes for child speech transformer-based phone recognition," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2021.

[146] D. Elenius and M. Blomberg, "Adaptation and normalization experiments in speech recognition for 4 to 8 year old children." in *Interspeech*, 2005, pp. 2749–2752.

[147] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[148] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" 2014.

[149] D. C. Cireşan, U. Meier, and J. Schmidhuber, "Transfer learning for latin and chinese characters with deep neural networks," in *The 2012 International Joint Conference on Neural Networks (IJCNN)*, 2012, pp. 1–6.

[150] R. Takashima, T. Takiguchi, and Y. Ariki, "Two-step acoustic model adaptation for dysarthric speech recognition," in *ICASSP*, 2020, pp. 6104–6108.

[151] R. Tong, L. Wang, and B. Ma, "Transfer learning for children's speech recognition," *2017 International Conference on Asian Language Processing (IALP)*, pp. 36–39, 2017.

[152] Y. Zhang and Q. Yang, "An overview of multi-task learning," *National Science Review*, vol. 5, no. 1, pp. 30–43, 2018.

[153] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *ICML*, ser. ICML '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 160–167. [Online]. Available: https://doi.org/10.1145/1390156.1390177

[154] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[155] L. Xie, S. He, Z. Zhang, K. Lin, X. Bo, S. Yang, B. Feng, K. Wan, K. Yang, J. Yang *et al.*, "Domain-adversarial multi-task framework for novel therapeutic property prediction of compounds," *Bioinformatics*, vol. 36, no. 9, pp. 2848–2855, 2020.

[156] S. R. Madikeri, B. K. Khonglah, S. Tong, P. Motlicek, H. Bourlard, and D. Povey, "Lattice-free maximum mutual information training of multilingual speech recognition systems." in *INTERSPEECH*, 2020, pp. 4746–4750.

[157] A. Abad, P. Bell, A. Carmantini, and S. Renais, "Cross lingual transfer learning for zero-resource domain adaptation," in *ICASSP*, 2020, pp. 6909–6913.

[158] L. Wei, W. Dong, B. Lin, and J. Zhang, "Multi-task based mispronunciation detection of children speech using multi-lingual information," in *APSIPA ASC*. IEEE, 2019, pp. 1791–1794.

[159] G. Zavaliagkos and T. Colthurst, "Utilizing untranscribed training data to improve perfomance." in *LREC*. Citeseer, 1998, pp. 317–322.

[160] J. Ma, S. Matsoukas, O. Kimball, and R. Schwartz, "Unsupervised training on large amounts of broadcast news data," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 3. IEEE, 2006, pp. III–III.

[161] J. Sarzynska-Wawer, A. Wawer, A. Pawlak, J. Szymanowska, I. Stefaniak, M. Jarkiewicz, and L. Okruszek, "Detecting formal thought disorder by deep contextualized word representations," *Psychiatry Research*, vol. 304, p. 114135, 2021.

[162] O. Henaff, "Data-efficient image recognition with contrastive predictive coding," in *International conference on machine learning*. PMLR, 2020, pp. 4182–4192.

[163] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[164] 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," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[165] M. Riviere, A. Joulin, P.-E. Mazaré, and E. Dupoux, "Unsupervised pretraining transfers well across languages," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7414–7418.

[166] Y. Wang, J. Li, H. Wang, Y. Qian, C. Wang, and Y. Wu, "Wav2vec-switch: Contrastive learning from original-noisy speech pairs for robust speech recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7097–7101.

[167] J. Li, V. Manohar, P. Chitkara, A. Tjandra, M. Picheny, F. Zhang, X. Zhang, and Y. Saraf, "Accent-robust automatic speech recognition using supervised and unsupervised wav2vec embeddings," *arXiv preprint arXiv:2110.03520*, 2021.

[168] G. Xu, S. Yang, L. Ma, C. Li, and Z. Wu, "The tal system for the interspeech2021 shared task on automatic speech recognition for non-native childrens speech." in *Interspeech*, 2021, pp. 1294–1298.

[169] R. Jain, A. Barcovschi, M. 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*, 2023.

[170] R. Jain, A. Barcovschi, M. Yiwere, P. Corcoran, and H. Cucu, "Adaptation of whisper models to child speech recognition," *arXiv preprint arXiv:2307.13008*, 2023.

[171] R. Fan and A. Alwan, "Draft: A novel framework to reduce domain shifting in self-supervised learning and its application to children's asr," *arXiv preprint arXiv:2206.07931*, 2022.

[172] K. Demuth, J. Culbertson, and J. Alter, "Word-minimality, epenthesis and coda licensing in the early acquisition of english," *Language and speech*, vol. 49, no. 2, pp. 137–173, 2006.

[173] K. Demuth and A. Tremblay, "Prosodically-conditioned variability in children's production of french determiners," *Journal of child language*, vol. 35, no. 1, pp. 99–127, 2008.

[174] J. Gao, A. Li, and Z. Xiong, "Mandarin multimedia child speech corpus: Cass_child," in *2012 International Conference on Speech Database and Assessments*, 2012, pp. 7–12.

[175] K. Demuth, “The acquisition of sesotho,” in *The crosslinguistic study of language acquisition.* Psychology Press, 1992, pp. 557–638.

[176] P. B. Ramteke, S. Supanekar, P. Hegde, H. Nelson, V. Aithal, and S. Koolagudi, “Nitk kids’ speech corpus,” *emotion*, vol. 491, pp. 4–15, 2019.

[177] M. Garrote and A. Moreno Sandoval, “Chiede, a spontaneous child language corpus of spanish,” in *Proceedings of the 3rd International LABLITA Workshop in Corpus Linguistics*, 2008.

[178] S.-I. Ng, C. W.-Y. Ng, J. Wang, T. Lee, K. Y.-S. Lee, and M. C.-F. Tong, “Cuchild: A large-scale cantonese corpus of child speech for phonology and articulation assessment,” *arXiv preprint arXiv:2008.03188*, 2020.

[179] E. Lyakso, O. Frolova, E. Dmitrieva, A. Grigorev, H. Kaya, A. A. Salah, and A. Karpov, “Emochildru: emotional child russian speech corpus,” in *International Conference on Speech and Computer.* Springer, 2015, pp. 144–152.

[180] A. Hämäläinen, S. Rodrigues, A. Júdice, S. M. Silva, A. Calado, F. M. Pinto, and M. S. Dias, “The cng corpus of european portuguese children’s speech,” in *International Conference on Text, Speech and Dialogue.* Springer, 2013, pp. 544–551.

[181] G. Yeung, A. L. Bailey, A. Afshan, M. Tinkler, M. Q. Pérez, A. Martin, A. A. Pogossian, S. Spaulding, H. W. Park, M. Muco *et al.*, “A robotic interface for the administration of language, literacy, and speech pathology assessments for children.” in *SLaTE*, 2019, pp. 41–42.

[182] M. Russell, S. D’Arcy, M. Wong, A. Batliner, M. Blomberg, and M. Gerosa, “The pf-star children’s speech corpus,” in *Interspeech 2005*, 2005.

[183] F. Yu, Z. Yao, X. Wang, K. An, L. Xie, Z. Ou, B. Liu, X. Li, and G. Miao, “The slt 2021 children speech recognition challenge: Open datasets, rules and baselines,” in *2021 IEEE Spoken Language Technology Workshop (SLT).* IEEE, 2021, pp. 1117–1123.

[184] M. Russell, “The pf-star british english childrens speech corpus,” *The Speech Ark Limited*, 2006.

[185] E. Lyakso, O. Frolova, A. Kaliyev, V. Gorodnyi, A. Grigorev, and Y. Matveev, “Ad-child. ru: Speech corpus for russian children with atypical development,” in *International Conference on Speech and Computer.* Springer, 2019, pp. 299–308.

[186] A. Kazemzadeh, H. You, M. Iseli, B. Jones, X. Cui, M. Heritage, P. Price, E. Anderson, S. Narayanan, and A. Alwan, “Tball data collection: the making of a young children’s speech corpus,” in *Ninth European Conference on Speech Communication and Technology*, 2005.

[187] F. Csatári, Z. Bakcsi, and K. Vicsi, "A hungarian child database for speech processing applications," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[188] A. Eshky, M. S. Ribeiro, J. Cleland, K. Richmond, Z. Roxburgh, J. Scobbie, and A. Wrench, "Ultrasuite: a repository of ultrasound and acoustic data from child speech therapy sessions," *arXiv preprint arXiv:1907.00835*, 2019.

[189] S. Lee, A. Potamianos, and S. 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.

[190] M. Khanzadi, H. Veisi, R. Alinaghizade, and Z. Soleymani, "Persian phoneme and syllable recognition using recurrent neural networks for phonological awareness assessment," *Journal of AI and Data Mining*, vol. 10, no. 1, pp. 117–126, 2022.

[191] J. Proença, D. Celorico, S. Candeias, C. Lopes, and F. Perdigão, "The letsread corpus of portuguese children reading aloud for performance evaluation," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 781–785.

[192] R. M. Pascual and R. C. L. Guevara, "Developing a children's filipino speech corpus for application in automatic detection of reading miscues and disfluencies," in *TENCON 2012 IEEE Region 10 Conference*, 2012, pp. 1–6.

[193] H. Pérez-Espinosa, J. Martínez-Miranda, I. Espinosa-Curiel, J. Rodríguez-Jacobo, L. Villaseñor-Pineda, and H. Avila-George, "Iesc-child: An interactive emotional children's speech corpus," *Computer Speech & Language*, vol. 59, pp. 55–74, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230817301547

[194] A. Hagen, B. Pellom, and R. Cole, "Children's speech recognition with application to interactive books and tutors," in *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721)*. IEEE, 2003, pp. 186–191.

[195] L. Cleuren, J. Duchateau, P. Ghesquiere *et al.*, "Children's oral reading corpus (chorec): description and assessment of annotator agreement," *LREC 2008 Proceedings*, pp. 998–1005, 2008.

[196] P. COSI, G. PACI, G. SOMMAVILLA, and F. TESSER, "Childit2–a new children read speech corpus."

[197] R. G. Leonard and G. Doddington, "Tidigits speech corpus," *Texas Instruments, Inc*, 1993.

[198] M. Gerosa, "Acoustic modeling for automatic recognition of children's speech," Ph.D. dissertation, Ph. D. thesis, University of Trento, 2006.

[199] F. Burkhardt, M. Eckert, W. Johannsen, and J. Stegmann, "A database of age and gender annotated telephone speech," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010.

[200] C. Cucchiarini, J. Driesen, H. Van hamme, and E. Sanders, "Recording speech of children, non-natives and elderly people for HLT applications: the JASMIN-CGN corpus." in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA), May 2008. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2008/pdf/366_paper.pdf

[201] L. Bell, J. Boye, J. Gustafson, M. Heldner, A. Lindström, and M. Wirén, "The swedish nice corpus–spoken dialogues between children and embodied characters in a computer game scenario," in *Interspeech 2005-Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*. ISCA, 2005, pp. 2765–2768.

[202] D. J. Iskra, B. Grosskopf, K. Marasek, H. van den Heuvel, F. Diehl, and A. Kiessling, "Speecon – speech databases for consumer devices: Database specification and validation," in *LREC*, 2002.

[203] R. Gretter, M. Matassoni, S. Bannò, and D. Falavigna, "Tlt-school: a corpus of non native children speech," 2020.

[204] H. Grissemann and M. Linder, "Zürcher lesetest," *Bern: Huber Verlag*, 2000.

[205] S. Steidl, *Automatic classification of emotion related user states in spontaneous children's speech.* Logos-Verlag Berlin, Germany, 2009.

[206] L. Bell and J. Gustafson, "Child and adult speaker adaptation during error resolution in a publicly available spoken dialogue system," in *Eighth European Conference on Speech Communication and Technology*. Citeseer, 2003.

[207] R. Nisimura, A. Lee, H. Saruwatari, and K. Shikano, "Public speech-oriented guidance system with adult and child discrimination capability," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2004, pp. I–433.

[208] M. Rayner, N. Tsourakis, C. Baur, P. Bouillon, and J. Gerlach, "Call-slt: A spoken call system: based on grammar and speech recognition," *Linguistic Issues in Language Technology*, vol. 10, 01 2014.

[209] X. Huang, F. Alleva, M.-Y. Hwang, and R. Rosenfeld, "An overview of the sphinx-ii speech recognition system," in *Proceedings of the Workshop on Human Language Technology*, ser. HLT '93. USA: Association for Computational Linguistics, 1993, p. 81–86. [Online]. Available: https://doi.org/10.3115/1075671.1075690

[210] V. Bhardwaj, V. Kukreja, Y. Belkhier, M. Bajaj, S. G. .B, A. Rehman, H. Hamam, and M. Othman, "Automatic speech recognition (asr) system for children's: A systematic literature review," *Applied Sciences*, 04 2022.

[211] C. F. Carvalho and A. Abad, "Tribus: An end-to-end automatic speech recognition system for european portuguese," *IberSPEECH 2021*, 2021.

[212] J. P. Neto, C. A. Martins, H. Meinedo, and L. B. Almeida, "The design of a large vocabulary speech corpus for portuguese," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[213] K. Paliwal, "Spectral subband centroid features for speech recognition," in *ICASSP*, vol. 2, 1998, pp. 617–620 vol.2.

[214] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv*, vol. abs/2111.09296, 2021.

[215] A. Zeyer, P. Bahar, K. Irie, R. Schlüter, and H. Ney, "A comparison of transformer and lstm encoder decoder models for asr," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 8–15.

[216] M. Zeineldeen, J. Xu, C. Lüscher, W. Michel, A. Gerstenberger, R. Schlüter, and H. Ney, "Conformer-based hybrid asr system for switchboard dataset," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7437–7441.

[217] C. Lüscher, E. Beck, K. Irie, M. Kitza, W. Michel, A. Zeyer, R. Schlüter, and H. Ney, "Rwth asr systems for librispeech: Hybrid vs attention–w/o data augmentation," *arXiv preprint arXiv:1905.03072*, 2019.

[218] H. Soltau, H. Liao, and H. Sak, "Neural speech recognizer: Acoustic-to-word lstm model for large vocabulary speech recognition," *arXiv preprint arXiv:1610.09975*, 2016.

[219] E. Battenberg, J. Chen, R. Child, A. Coates, Y. G. Y. Li, H. Liu, S. Satheesh, A. Sriram, and Z. Zhu, "Exploring neural transducers for end-to-end speech recognition," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 206–213.

[220] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[221] A. Bahrini, M. Khamoshifar, H. Abbasimehr, R. J. Riggs, M. Esmaeili, R. M. Majdabadkohne, and M. Pasehvar, "Chatgpt: Applications, opportunities, and threats," in *2023 Systems and Information Engineering Design Symposium (SIEDS)*. IEEE, 2023, pp. 274–279.

[222] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8821–8831.

[223] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, "Attention augmented convolutional networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3286–3295.

[224] B. Yang, L. Wang, D. Wong, L. S. Chao, and Z. Tu, "Convolutional self-attention networks," *arXiv preprint arXiv:1904.03107*, 2019.

[225] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[226] Y. Lu, Z. Li, D. He, Z. Sun, B. Dong, T. Qin, L. Wang, and T.-Y. Liu, "Understanding and improving transformer from a multi-particle dynamic system point of view," *arXiv preprint arXiv:1906.02762*, 2019.

[227] Z. Wu, Z. Liu, J. Lin, Y. Lin, and S. Han, "Lite transformer with long-short range attention," *arXiv preprint arXiv:2004.11886*, 2020.

[228] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *International conference on machine learning*. PMLR, 2017, pp. 933–941.

[229] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," *CoRR*, vol. abs/1710.05941, 2017. [Online]. Available: http://arxiv.org/abs/1710.05941

[230] O. Kovaleva, A. Romanov, A. Rogers, and A. Rumshisky, "Revealing the dark secrets of BERT," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 4365–4374. [Online]. Available: https://aclanthology.org/D19-1445

[231] P. Michel, O. Levy, and G. Neubig, "Are sixteen heads really better than one?" *Advances in neural information processing systems*, vol. 32, 2019.

[232] P. Ye, Y. Huang, C. Tu, M. Li, T. Chen, T. He, and W. Ouyang, "Partial fine-tuning: A successor to full fine-tuning for vision transformers," 2023.

[233] J. McCarley, R. Chakravarti, and A. Sil, "Structured pruning of a bert-based question answering model," *arXiv preprint arXiv:1910.06360*, 2019.

[234] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[235] Z. Shen, Z. Liu, J. Qin, M. Savvides, and K.-T. Cheng, "Partial is better than all: revisiting fine-tuning strategy for few-shot learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 11, 2021, pp. 9594–9602.

[236] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021.

[237] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," *arXiv preprint arXiv:1803.03635*, 2018.

[238] S. Gandhi, P. von Platen, and A. M. Rush, "Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling," 2023.

[239] H.-J. Chang, S.-w. Yang, and H.-y. Lee, "Distilhubert: Speech representation learning by layer-wise distillation of hidden-unit bert," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7087–7091.

[240] Y. Peng, Y. Sudo, S. Muhammad, and S. Watanabe, "DPHuBERT: Joint Distillation and Pruning of Self-Supervised Speech Models," in *Proc. INTERSPEECH 2023*, 2023, pp. 62–66.

[241] 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.

[242] M. Geva, R. Schuster, J. Berant, and O. Levy, "Transformer feed-forward layers are key-value memories," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[243] J. Kaplan, S. McCandlish, T. J. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *ArXiv*, vol. abs/2001.08361, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:210861095

[244] W. Zheng, A. Xiao, G. Keren, D. Le, F. Zhang, C. Fuegen, O. Kalinli, Y. Saraf, and A. Mohamed, "Scaling ASR Improves Zero and Few Shot Learning," in *Proc. Interspeech 2022*, 2022, pp. 5135–5139.

[245] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2790–2799.

[246] J. Pfeiffer, I. Vulić, I. Gurevych, and S. Ruder, "MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 7654–7673.

[247] A. Rücklé, G. Geigle, M. Glockner, T. Beck, J. Pfeiffer, N. Reimers, and I. Gurevych, "AdapterDrop: On the efficiency of adapters in transformers," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 7930–7946. [Online]. Available: https://aclanthology.org/2021.emnlp-main.626

[248] U. Cappellazzo, D. Falavigna, A. Brutti, and M. Ravanelli, "Parameter-efficient transfer learning of audio spectrogram transformers," *arXiv preprint arXiv:2312.03694*, 2023.

[249] N. Chen, I. Shafran, Y. Zhang, C.-C. Chiu, H. Soltau, J. Qin, and Y. Wu, "Efficient adapters for giant speech models," *arXiv preprint arXiv:2306.08131*, 2023.

[250] S. V. Eeckt and H. Van Hamme, "Using adapters to overcome catastrophic forgetting in end-to-end automatic speech recognition," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[251] A. Kannan, A. Datta, T. N. Sainath, E. Weinstein, B. Ramabhadran, Y. Wu, A. Bapna, Z. Chen, and S. Lee, "Large-scale multilingual speech recognition with a streaming end-to-end model," *arXiv preprint arXiv:1909.05330*, 2019.

[252] W. Hou, H. Zhu, Y. Wang, J. Wang, T. Qin, R. Xu, and T. Shinozaki, "Exploiting adapters for cross-lingual low-resource speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 317–329, 2021.

[253] A. Kulkarni, A. Kulkarni, M. Couceiro, and H. Aldarmaki, "Adapting the adapters for code-switching in multilingual asr," *arXiv preprint arXiv:2310.07423*, 2023.

[254] B. Thomas, S. Kessler, and S. Karout, "Efficient adapter transfer of self-supervised speech models for automatic speech recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7102–7106.

[255] K. Tomanek, V. Zayats, D. Padfield, K. Vaillancourt, and F. Biadsy, "Residual adapters for parameter-efficient asr adaptation to atypical and accented speech," *arXiv preprint arXiv:2109.06952*, 2021.

[256] J. He, C. Zhou, X. Ma, T. Berg-Kirkpatrick, and G. Neubig, "Towards a unified view of parameter-efficient transfer learning," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=0RDcd5Axok

[257] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[258] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.

[259] A. Fazel, W. Yang, Y. Liu, R. Barra-Chicote, Y. Meng, R. Maas, and J. Droppo, "SynthASR: Unlocking Synthetic Data for Speech Recognition," in *Proc. Interspeech 2021*, 2021, pp. 896–900.

[260] T.-Y. Hu, M. Armandpour, A. Shrivastava, J.-H. R. Chang, H. Koppula, and O. Tuzel, "Synt++: Utilizing imperfect synthetic data to improve speech recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7682–7686.

[261] E. Casanova, C. Shulby, A. Korolev, A. C. Junior, A. da Silva Soares, S. Aluísio, and M. A. Ponti, "ASR data augmentation in low-resource settings using cross-lingual multi-speaker TTS and cross-lingual voice conversion," in *Proc. INTERSPEECH 2023*, 2023, pp. 1244–1248.

[262] S. Ueno, M. Mimura, S. Sakai, and T. Kawahara, "Data augmentation for asr using tts via a discrete representation," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 68–75.

[263] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=rylwJxrYDS

[264] J. Philip, A. Berard, M. Gallé, and L. Besacier, "Monolingual adapters for zero-shot neural machine translation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 4465–4470.

[265] Y. Mao, L. Mathias, R. Hou, A. Almahairi, H. Ma, J. Han, S. Yih, and M. Khabsa, "UniPELT: A unified framework for parameter-efficient language model tuning," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 6253–6264. [Online]. Available: https://aclanthology.org/2022.acl-long.433

[266] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17–23 Jul 2022, pp. 2709–2720. [Online]. Available: https://proceedings.mlr.press/v162/casanova22a.html

[267] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real nvp," 2017. [Online]. Available: https://arxiv.org/abs/1605.08803

[268] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," in *Arxiv*, 2016. [Online]. Available: https://arxiv.org/abs/1609.03499

[269] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[270] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *CoRR*, vol. abs/1312.6114, 2014.

[271] H. S. Heo, B.-J. Lee, J. Huh, and J. S. Chung, "Clova baseline system for the voxceleb speaker recognition challenge 2020," 2020.

[272] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2016.

[273] E. Casanova, A. C. Junior, C. Shulby, F. S. d. Oliveira, J. P. Teixeira, M. A. Ponti, and S. Aluísio, "Tts-portuguese corpus: a corpus for speech synthesis in brazilian portuguese," *Language Resources and Evaluation*, vol. 56, no. 3, pp. 1043–1055, 2022.

[274] Munich Artificial Intelligence Laboratories GmbH, "The mailabs speech dataset – caito," 2017, accessed January 15, 2024. https://www.caito.de/2019/01/03/the-m-ailabs-speech-dataset/.

[275] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech 2020*, 2020, pp. 3830–3834.

[276] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9

[277] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3045–3059. [Online]. Available: https://aclanthology.org/2021.emnlp-main.243

[278] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 4582–4597. [Online]. Available: https://aclanthology.org/2021.acl-long.353

[279] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11 966–11 976.

[280] L.-J. Yang, C.-H. H. Yang, and J.-T. Chien, "Parameter-Efficient Learning for Text-to-Speech Accent Adaptation," in *Proc. INTERSPEECH 2023*, 2023, pp. 4354–4358.

[281] N. Muthuchamy Selvaraj, X. Guo, A. Kong, B. Shen, and A. Kot, "Adapter Incremental Continual Learning of Efficient Audio Spectrogram Transformers," in *Proc. INTERSPEECH 2023*, 2023, pp. 909–913.

[282] S. Jie and Z.-H. Deng, "Convolutional bypasses are better vision transformer adapters," *arXiv preprint arXiv:2207.07039*, 2022.

[283] Y. Li, A. Mehrish, S. Zhao, R. Bhardwaj, A. Zadeh, N. Majumder, R. Mihalcea, and S. Poria, "Evaluating parameter-efficient transfer learning approaches on sure benchmark for speech understanding," in *ICASSP*, 2023.

[284] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[285] E. Ben Zaken, Y. Goldberg, and S. Ravfogel, "BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 1–9. [Online]. Available: https://aclanthology.org/2022.acl-short.1

[286] D. Lian, D. Zhou, J. Feng, and X. Wang, "Scaling & shifting your features: A new baseline for efficient model tuning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 109–123, 2022.

[287] Y. Wu and K. He, "Group normalization," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[288] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1501–1510.

[289] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.

[290] C.-L. Fu, Z.-C. Chen, Y.-R. Lee, and H.-y. Lee, "AdapterBias: Parameter-efficient token-dependent representation shift for adapters in NLP tasks," in *Findings of the Association for Computational Linguistics: NAACL 2022*, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 2608–2621. [Online]. Available: https://aclanthology.org/2022.findings-naacl.199

[291] T. P. Pires, A. V. Lopes, Y. Assogba, and H. Setiawan, "One wide feedforward is all you need," *arXiv preprint arXiv:2309.01826*, 2023.

[292] Y. Hauptman, R. Aloni-Lavi, I. Lapidot, T. Gurevich, Y. Manor, S. Naor, N. Diamant, and I. Opher, "Identifying distinctive acoustic and spectral features in parkinson's disease." in *Interspeech*, 2019, pp. 2498–2502.

[293] M. C. Botelho, I. Trancoso, A. Abad, and T. Paiva, "Speech as a biomarker for obstructive sleep apnea detection," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5851–5855.

[294] I. Laaridh, W. B. Kheder, C. Fredouille, and C. Meunier, "Automatic Prediction of Speech Evaluation Metrics for Dysarthric Speech," in *Proc. Interspeech 2017*, 2017, pp. 1834–1838.

[295] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "x-vectors meet emotions: A study on dependencies between emotion and speaker recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7169–7173.

[296] J. M. Perero-Codosero, F. Espinoza-Cuadros, J. Antón-Martín, M. A. Barbero-Alvarez, and L. A. Hernández-Gómez, "Modeling obstructive sleep apnea voices using deep neural network embeddings and domain-adversarial training," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 240–250, 2019.

[297] S. Zargarbashi and B. Babaali, "A multi-modal feature embedding approach to diagnose alzheimer disease from spoken language," *arXiv preprint arXiv:1910.00330*, 2019.

[298] C. Botelho, F. Teixeira, T. Rolland, A. Abad, and I. Trancoso, "Pathological speech detection using x-vector embeddings," *arXiv preprint arXiv:2003.00864*, 2020.

[299] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification." in *Interspeech*, vol. 2017, 2017, pp. 999–1003.

[300] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[301] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[302] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[303] A. Hämäläinen, J. Avelar, S. Rodrigues, J. Dias, A. Kolesinski, T. Fegyó, G. Németh, P. Csobánka, K. Ting, and D. Hewson, "The easr corpora of european portuguese, french, hungarian and polish elderly speech," *The EASR Corpora of European Portuguese, French, Hungarian and Polish elderly speech*, pp. 1458–1464, 2014.

[304] S. Pinto, R. Cardoso, J. Sadat, I. Guimarães, C. Mercier, H. Santos, C. Atkinson-Clement, J. Carvalho, P. Welby, P. Oliveira *et al.*, "Dysarthria in individuals with parkinson's disease: a protocol for a binational, cross-sectional, case-controlled study in french and european portuguese (fralusopark)," *BMJ open*, vol. 6, no. 11, p. e012885, 2016.

[305] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, M. C. Gonzalez-Rátiva, and E. Nöth, "New spanish speech corpus database for the analysis of people suffering from parkinson's disease." in *LREC*, 2014, pp. 342–347.

[306] A. Pompili, A. Abad, P. Romano, I. P. Martins, R. Cardoso, H. Santos, J. Carvalho, I. Guimaraes, and J. J. Ferreira, "Automatic detection of parkinson's disease: an experimental analysis of common

speech production tasks used for diagnosis," in *International Conference on Text, Speech, and Dialogue*. Springer, 2017, pp. 411–419.

[307] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[308] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 835–838.

[309] A. Pompili, T. Rolland, and A. Abad, "The inesc-id multi-modal system for the adress 2020 challenge," *arXiv preprint arXiv:2005.14646*, 2020.

[310] A. König, A. Satt, A. Sorin, R. Hoory, O. Toledo-Ronen, A. Derreumaux, V. Manera, F. Verhey, P. Aalten, P. H. Robert *et al.*, "Automatic speech analysis for the assessment of patients with predementia and alzheimer's disease," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 1, no. 1, pp. 112–124, 2015.

[311] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2016.

[312] G. Gosztolya, V. Vincze, L. Tóth, M. Pákáski, J. Kálmán, and I. Hoffmann, "Identifying mild cognitive impairment and mild alzheimer's disease based on spontaneous speech using asr and linguistic features," *Computer Speech & Language*, vol. 53, pp. 181–197, 2019.

[313] T. Warnita, N. Inoue, and K. Shinoda, "Detecting Alzheimer's Disease Using Gated Convolutional Neural Network from Audio Data," in *Proc. Interspeech 2018*, 2018, pp. 1706–1710.

[314] S. Karlekar, T. Niu, and M. Bansal, "Detecting linguistic characteristics of Alzheimer's dementia by interpreting neural models," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, M. Walker, H. Ji, and A. Stent, Eds. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 701–707. [Online]. Available: https://aclanthology.org/N18-2110

[315] H. Goodglass, E. Kaplan, and S. Weintraub, *BDAE: The Boston diagnostic aphasia examination.* Lippincott Williams & Wilkins Philadelphia, PA, 2001.

[316] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The adress challenge," *arXiv preprint arXiv:2004.06833*, 2020.

[317] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Proc. Interspeech 2017*, 2017, pp. 2616–2620.

[318] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen *et al.*, "The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly Emotion, Breathing & Masks," *Proceedings INTERSPEECH. Shanghai, China: ISCA*, 2020.

[319] R. Solera-Ureña, C. Botelho, F. Teixeira, T. Rolland, A. Abad, and I. Trancoso, "Transfer learning-based cough representations for automatic detection of covid-19," in *Interspeech*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:237489901

[320] B. W. Schuller, A. Batliner, C. Bergler, C. Mascolo, J. Han, I. Lefter, H. Kaya, S. Amiriparian, A. Baird, L. Stappen, S. Ottl, M. Gerczuk, P. Tzirakis, C. Brown, J. Chauhan, A. Grammenos, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, L. J. Rothkrantz, J. A. Zwerts, J. Treep, and C. S. Kaandorp, "The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation and Primates," in *Proc. Interspeech 2021*, 2021, pp. 431–435.

[321] R. X. A. Pramono, S. A. Imtiaz, and E. Rodriguez-Villegas, "A cough-based algorithm for automatic diagnosis of pertussis," *PloS one*, vol. 11, no. 9, p. e0162128, 2016.

[322] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 131–135.

[323] C. Brown, J. Chauhan, A. Grammenos, J. Han, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, and C. Mascolo, "Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Virtual Event, CA, USA, 2020, p. 3474–3484.

[324] P. Bagad, A. Dalmia, J. Doshi, A. Nagrani, P. Bhamare, A. Mahale, S. Rane, N. Agarwal, and R. Panicker, "Cough Against COVID: Evidence of COVID-19 Signature in Cough Sounds," *preprint arXiv:2009.08790*, 2020.

[325] A. Imran, I. Posokhova, H. N. Qureshi, U. Masood, M. S. Riaz, K. Ali, C. N. John, M. I. Hussain, and M. Nabeel, "AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app," *Informatics in Medicine Unlocked*, vol. 20, p. 100378, 2020.

[326] G. Chaudhari, X. Jiang, A. Fakhry, A. Han, J. Xiao, S. Shen, and A. Khanzada, "Virufy: Global Applicability of Crowdsourced and Clinical Datasets for AI Detection of COVID-19 from Cough," *preprint arXiv:2011.13320*, 2021.

[327] J. Han, C. Brown, J. Chauhan, A. Grammenos, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, and C. Mascolo, "Exploring Automatic COVID-19 Diagnosis via Voice and Symptoms from Crowdsourced Data," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 8328–8332.

[328] L. Orlandic, T. Teijeiro, and D. Atienza, "The COUGHVID crowdsourcing dataset: A corpus for the study of large-scale cough analysis algorithms," *preprint arXiv:2009.11644*, 2020.

[329] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, L. P. García-Perera, F. Richardson, R. Dehak, P. A. Torres-Carrasquillo, and N. Dehak, "State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and Speakers in the Wild evaluations," *Computer Speech & Language*, vol. 60, p. 101026, 2020.

[330] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *3rd International Conference on Learning Representations, ICLR 2015*, San Diego, CA, USA, 2015.

[331] S. Pascual, M. Ravanelli, J. Serrà, A. Bonafonte, and Y. Bengio, "Learning Problem-Agnostic Speech Representations from Multiple Self-Supervised Tasks," in *Proceedings of Interspeech 2019*, Graz, Austria, 2019, pp. 161–165.

[332] M. Ravanelli, J. Zhong, S. Pascual, P. Swietojanski, J. Monteiro, J. Trmal, and Y. Bengio, "Multi-Task Self-Supervised Learning for Robust Speech Recognition," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6989–6993.

[333] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech 2021*, 2021, pp. 1194–1198.

[334] X. Chang, T. Maekaku, P. Guo, J. Shi, Y.-J. Lu, A. S. Subramanian, T. Wang, S.-w. Yang, Y. Tsao, H.-y. Lee *et al.*, "An exploration of self-supervised pretrained representations for end-to-end speech recognition," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 228–235.

[335] A. T. Liu, S.-w. Yang, P.-H. Chi, P.-c. Hsu, and H.-y. Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," *ICASSP 2020 - 2020*

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2020. [Online]. Available: http://dx.doi.org/10.1109/ICASSP40776.2020.9054458

[336] P.-H. Chi, P.-H. Chung, T.-H. Wu, C.-C. Hsieh, Y.-H. Chen, S.-W. Li, and H.-y. Lee, "Audio albert: A lite bert for self-supervised learning of audio representation," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 344–350.

[337] A. H. Liu, Y.-A. Chung, and J. Glass, "Non-Autoregressive Predictive Coding for Learning Speech Representations from Local Dependencies," in *Proc. Interspeech 2021*, 2021, pp. 3730–3734.

[338] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. Glass, "An Unsupervised Autoregressive Model for Speech Representation Learning," in *Proc. Interspeech 2019*, 2019, pp. 146–150.

[339] A. T. Liu, S.-W. Li, and H.-y. Lee, "Tera: Self-supervised learning of transformer encoder representation for speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2351–2366, 2021.

[340] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazaré, 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)*, 2020, pp. 7669–7673, https://github.com/facebookresearch/libri-light.

[341] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.

[342] L. Gelin, "Reconnaissance automatique de la parole d'enfants apprenant·e·s lecteur·ice·s en salle de classe : modélisation acoustique de phonèmes," Ph.D. dissertation, 02 2022.

A

Pathological speech detection through pre-trained models

A.1 Introduction

In the domain of speech therapy, and specifically paediatric speech therapy, advancements in speech and language technologies hold significant promise by providing automated tools for assessing pronunciation quality and identifying pathological conditions. While the primary focus of this thesis was to improve ASR for children, we also contributed into the identification of pathological conditions from speech. This annex provides a overview of these contributions.

A.2 Pathological speech detection using x-vector embeddings

A.2.1 Introduction

Speech has been proposed as a valuable biomarker for detecting various diseases, including neurological conditions, mood disorders, and respiratory diseases [292, 293]. However, challenges such as temporal and financial constraints, lack of medical community awareness, ethical concerns, and patient privacy laws impede the acquisition of medical data, posing significant obstacles to the development of health-related speech-based classifiers, especially for deep learning models.

Most existing systems rely on Knowledge-Based (KB) features, often limited in capturing subtle symptoms and variations in disease severity. To address this limitation, some studies focus on speaker representation models, such as Gaussian Supervectors and i-vectors. For instance, [292] proposed i-vectors for Parkinson's disease classification, while [294] applied the i-vector paradigm to predict dysarthric speech evaluation metrics. The rationale behind using these representations lies in their ability to model speaker variability, which should also including disease symptoms [292].

X-vectors are discriminative DNN-based speaker embeddings, surpassing i-vectors in tasks like speaker and language recognition [257]. Despite initial doubts about the usability of such discriminative representations for disease detection as they have been trained on general datasets without diseased patients, recent studies have demonstrated their effectiveness. X-vectors have been successfully applied to paralinguistic tasks such as emotion recognition [295], obstructive sleep apnea detection [296], and as a complement to Alzheimer's Disease detection [297]. In our work [298], we investigate the hypothesis that speaker characteristics embedded in x-vectors, obtained from a single network trained for speaker identification using general data, contain sufficient information for the detection of multiple diseases. Furthermore, we aim to assess whether this information persists even in the presence of language mismatch, a phenomenon previously observed in speaker recognition [299]. Specifically, we employ the x-vector model as a feature extractor to train Support Vector Machine (SVM) for detecting two speech-affecting diseases: Parkinson's disease (PD) and Obstructive Sleep Apnea (OSA).

A.2.2 Speaker embeddings: i-vector and x-vector

Speaker embeddings serve as fixed-length representations of variable-length speech signals, capturing essential information about the speaker. Traditional methods, such as Gaussian Supervectors [300] derived from MAP-adapted Gaussian Mixture Model - Universal Background Model (GMM-UBM) [301] and i-vectors [302], have been fundamental in speaker recognition.

I-vectors, until recently considered state-of-the-art, extend the GMM Supervector approach by modeling total variability as a low-rank space through factor analysis. [292] observed that i-vectors, capturing total variability and speaker variability, also encompass information about speech disorders. For classification, they used reference i-vectors for healthy and PD populations.

In contrast, x-vectors, proposed as an alternative to i-vectors, aim to discriminate between speakers by modeling specific characteristics. Unlike i-vectors, x-vectors exhibit robustness to data variability and domain mismatches, requiring shorter temporal segments for optimal performance. Typically, the x-vector system comprises three main blocks: TDNN layers operating at the frame level, a statistical pooling layer for temporal aggregation (employing an attentive mechanism for importance weighting), and fully connected (Dense) layers for x-vector extraction.

A.2.3 Experimental setup

In our experiments, we used four corpora to determine the presence or absence of PD and OSA. With one of the European Portuguese corpus was employed to train the i-vector and x-vector extractors. For each disease-related dataset, we compared three representations: KB features, i-vectors, and x-vectors. All disease classifications were conducted using a SVM classifier using leave-one-speaker-out cross validation as an alternative to partitioning the corpora into train, development and test sets. Further details on the corpora, data representations, and classification method are provided below.

Our classification process operates at the segment level, assigning speakers a final classification through a weighted majority voting mechanism. In this approach, predictions obtained for each segment uttered by the speaker are weighted based on the corresponding number of speech frames.

A.2.3.A Corpora

In this section, we provide a description of the datasets employed in our study. The Speaker Recognition - Portuguese (PT-EASR) Corpus is a subset of the EASR (Elderly Automatic Speech Recognition) corpus [303]. It comprises recordings of European Portuguese read sentences and was used for training both i-vector and x-vector models for speaker recognition tasks. The dataset encompasses speakers aged 24 to 91, with 91% falling within the 60-80 age range. This specific age distribution was chosen with the intention of generating reliable speaker embeddings for this age group, particularly relevant to the diseases addressed in our study. The corpus was partitioned into training, development, and test sets in a ratio of 0.70:0.15:0.15, respectively. For PD Detection - Portuguese PD (PPD) Corpus, a subset of the FraLusoPark corpus [304] was employed. This subset includes speech recordings of both French and European Portuguese healthy volunteers and PD patients. The selected utterances consist of European Portuguese speakers reading prosodic sentences. The PD Detection - Spanish PD (SPD) Corpus corresponds to a subset of the New Spanish Parkinson's Disease Corpus, collected at the Universidad de Antioquia, Colombia [305]. For this experiment, we only used the read sentences subset. This corpus serves the purpose of investigating whether x-vector representations trained in one language (European Portuguese) can generalise effectively to another language, namely Spanish. The

OSA Detection - PSD Corpus is an extended version of the Portuguese Sleep Disorders (PSD) corpus [293]. This corpus introduces tasks in European Portuguese, including reading a phonetically rich text, read sentences recorded during a cognitive load assessment task, and spontaneous descriptions of an image. All utterances were segmented into 4-second-long segments using overlapping windows with a 2-second shift. Further details about each of these datasets can be found in Table A.1.

Language	Task	Group	Speakers	Segments	Duration (h)
PT	Spk. Rcg.	-	919	290,690	171.81
	PD	Patient	75	1,838	1.24
		Control	65	1,527	1.07
	OSA	Patient	30	1,793	1.10
		Control	30	1,702	1.05
SP	PD	Patient	50	661	0.49
		Control	50	655	0.50

Table A.1: Description of Speakers and Segments

A.2.3.B Knowledge based features

For PD classification, the KB feature set proposed by Pompili et al. [306] comprises 36 features from the eGeMAPS [307], along with the mean and standard deviation (std) of 12 MFCCs and log-energy. Additionally, the set includes the first and second derivatives of these coefficients, resulting in a 114-dimensional feature vector.

In the case of OSA classification, the KB feature set, as proposed in [293], includes the mean of 12 MFCCs, along with their first and second order derivatives, and 48 linear prediction cepstral coefficients. The set also covers the mean and std of the frequency and bandwidth of formant 1, 2, and 3, as well as the mean and std of Harmonics-to-noise ratio, jitter, F0 at percentiles 20, 50, and 100, and mean and std values for all frames and only voiced frames of Spectral Flux. All KB features were extracted using openSMILE [308].

A.2.3.C Speaker embeddings

In the i-vector system, 19 MFCCs plus log-energy are given as input with non-speech frames removed using energy-based Voice Activity Detection (VAD). Utterances are modeled with a 512-component full-covariance GMM, resulting in 180-dimensional i-vectors. The entire process is implemented using Kaldi [69] over the PT-EASR corpus.

For x-vectors, the network architecture is detailed in Table A.2, and x-vectors are extracted at the 6^{th} layer (Dense 6). Using 24-dimensional fbanks as input features. The non-speech frames were removed using energy-based VAD. This network was trained on the PT-EASR corpus for speaker identification, using 100 epochs, cross-entropy loss, a learning rate of 0.001, a learning rate decay of 0.05 with a 30-epoch

Layer	Contex	Total Contex	In $\times$ Out
TDNN1	$[t-2, t+2]$	5	$5F \times 256$
TDNN2	$\{t-2, t, t+2\}$	9	768×256
TDNN3	$\{t-3, t, t+3\}$	15	768×256
TDNN4	$\{t\}$	15	256×256
TDNN5	$\{t\}$	15	256×512
stats pooling	$[0, T)$	T	$512T \times 1024$
Dense 6	$\{0\}$	T	1024×512
Dense 7	$\{0\}$	T	512×512
softmax	$\{0\}$	T	$512 \times S$

Table A.2: X-vector network Description

period, a batch size of 512, and a dropout value of 0.001.

A.2.4 Results

		PD - Portuguese			OSA			PD - Spanish		
Features		Prec.	Recall	F1 Score	Prec.	Recall	F1 Score	Prec.	Recall	F1 Score
KB	Seg	64.5	64.6	64.5	64.8	64.9	64.8	**79.0**	**79.0**	**79.0**
	Spk	72.2	72.3	72.1	**82.0**	**81.7**	81.6	**87.1**	**87.0**	**87.0**
i-vector	Seg	66.6	66.6	66.6	65.6	65.6	65.6	75.7	75.7	75.7
	Spk	**75.6**	**75.7**	**75.6**	72.3	75.0	75.0	85.1	85.0	85.0
x-vector	Seg	**66.7**	**66.8**	**66.7**	**73.3**	**73.3**	**73.3**	77.2	77.2	77.1
	Spk	74.4	74.5	74.3	81.7	**81.7**	**81.7**	86.0	86.0	86.0

Table A.3: Results of the different tasks with KB and speaker embeddings

The results of the different task are summarised in Table A.3. For PD with Portuguese data, the findings indicate that speaker representations learned from out-of-domain data surpass the performance of KB features. This supports our hypothesis that speaker embeddings not only capture information about speech pathologies but also model symptoms of the disease that KB features may fail to include.

It is notable that x-vectors and i-vectors yield very similar results, with a slight advantage for x-vectors at the segment level and slightly better results for i-vectors at the speaker level. This observation suggests that while x-vectors offer stronger representations for short segments, i-vectors may perform better for longer segments. The application of a majority vote weighted by the duration of speech segments may favor the i-vector approach at the speaker level.

In the context of the OSA task, x-vectors demonstrate superior performance compared to all other approaches at the segment level. Notably, they significantly showed around 8% improvement over KB features, providing further support for our hypothesis. However, it is essential to highlight that both x-vectors and i-vectors perform similarly at the speaker level. Interestingly, i-vectors, in this scenario, perform less effectively than KB features. One possible explanation could be attributed to the fact that the PSD corpus incorporates tasks, such as spontaneous speech, which diverge from the read sentences included in the corpus used to train the i-vector and x-vector extractors. These tasks might be considered out-of-domain, which would explain why x-vectors outperform the i-vector approach.

The objective of the PD Spanish experiment was to evaluate the efficacy of x-vectors trained in one language when applied to disease classification in a different language. Our findings reveal that KB features outperform both speaker representations, possibly due to the language mismatch between the Spanish PD corpus and the European Portuguese training corpus. However, it is noteworthy that, akin to the previous task, x-vectors demonstrate the ability to surpass i-vectors in an out-of-domain corpus.

To conclude, our experiments conducted on the European Portuguese datasets substantiate the hypothesis that speaker embeddings encompass pertinent information for disease detection. Notably, we identified evidence indicating that these embeddings capture information not represented by KB features, validating the efficacy of our approach. Additionally, the observations suggest that x-vectors outperform i-vectors in tasks where the domain does not align with the training data, such as verbal task mismatch and cross-lingual experiments. This underscores the potential of x-vector embeddings as formidable alternatives to KB feature sets for the detection of PD and OSA.

A.3 The INESC-ID Multi-Modal System for the ADReSS 2020 Challenge

A.3.1 Introduction

Later, in [309], we proposed to extend the aforementioned work by classifying Alzheimer's disease (AD). Indeed, existing studies have explored various approaches, including syntactic or semantic features, plain acoustic methods, and combinations of temporal speech parameters and lexical measures. However, the diversity in datasets and methodologies makes it challenging to compare these studies. To address this, the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) challenge was introduced, providing a common, statistically balanced, and acoustically enhanced dataset for researchers to test their approaches. In this work we introduced a the multi-modal system where both acoustic and textual feature embeddings were used for automatically distinguishing AD patients from healthy individual.

Initially, studies focused on hand-crafted temporal and acoustic parameters from speech or linguistic, or a fusion of both. For example, [310] analysed temporal speech features. [311] employed over 350 features to capture lexical, syntactic, grammatical, and semantic phenomena from transcriptions of a picture description task. Finally, [312] used in conjunction demographic, acoustic, and linguistic features.

More recently, a shift towards advanced architectures has been observed to overcome limitations in traditional methods. For instance, [313] used a gated CNN on acoustic data, while [314] explored linguistic impairments with CNN, RNNs, and a combination. Finally, [297] introduced a multi-modal

	Train		Test
	Control	AD	-
Audio Full	55min46s	1h14min	1h06min
Audio chunks	30min11s	26min31s	26min32s
# Words (unique)	6097 (567)	5494 (552)	5536 (602)

Table A.4: Statistical information on the ADReSS corpus

feature embedding approach, based on N-gram, i-vector and x-vectors.

Our work differs from previous studies by using contextual embedding vectors for text data, feeding into two systems: one employing Global Maximum pooling and bidirectional LSTM-RNNs architectures, and the other based on statistical computation of sentence embeddings. This approach is simpler and does not require training deep architectures. For audio, we use DNN speaker embeddings extracted from pre-trained models. This is the first work which jointly use automatically learned representations for both audio and textual data.

A.3.2 Corpus

The ADReSS dataset comprises speech recordings and annotated transcriptions from 156 subjects, including 78 AD patients and 78 age and gender-matched healthy controls. The data is split into training (108 subjects) and test (48 subjects) sets. Participants provided descriptions of the Cookie Theft picture from the Boston Diagnostic Aphasia Examination [315]. Speech recordings were segmentated using VAD and normalised [316]. The dataset contained both full enhanced audio, and normalised audio chunks. Our approach used both audio and transcriptions. The transcriptions were annotated with disfluencies, filled pauses, repetitions, and other complex events. The transcriptions contained 17,127 words, including 1,009 unique words. Additional details on the duration and size of the ADReSS dataset are provided in Table A.4.

A.3.3 Proposed system

Our multi-modal framework, illustrated in Figure A.1, is based on the independent generation of acoustic and textual feature embeddings. Subsequently, we conduct an early fusion of the output of the two systems to create a singular feature vector encapsulating a condensed representation of both speech and language characteristics. The final classification is carried out using an SVM classifier with a linear kernel. Further details on the two systems are provided in the subsequent sections.

A.3.3.A Acoustics modality

The acoustic system incorporates i-vectors and x-vectors. Taking into consideration the small size of the ADReSS dataset, we preferred to exploit already existing pre-trained models to produce our acoustic fea-

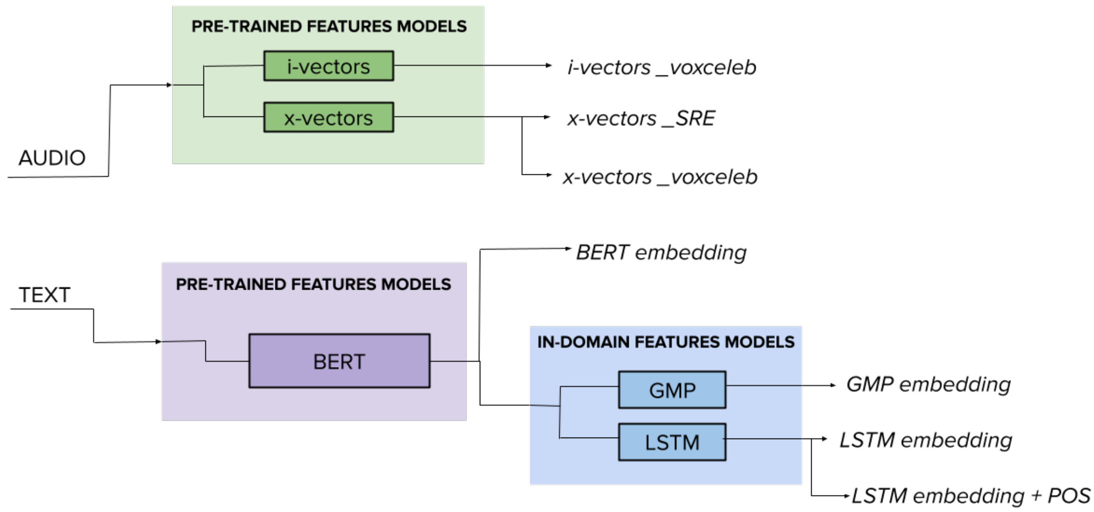


Figure A.1: Overview of the multimodal system based on embeddding appraoches

ture embeddings, rather than training them using in-domain challenge data. For the x-vectors framework, both the SRE and Voxceleb models were employed. The SRE model was primarily trained on telephone and microphone speech using data from the Switchboard corpus, Mixer 6, and NIST SREs [257]. The Voxceleb model was trained on augmented VoxCeleb 1 and VoxCeleb 2 datasets, encompassing speech from speakers with diverse ethnicities, accents, professions, and ages [257, 317]. This dataset was also used to build the i-vectors pre-trained model.

Inputs to these pre-trained models included 23 and 30-dimensional MFCCs extracted with Kaldi [69] and non-speech frames were filtered out using VAD. For x-vectors, a 512-dimensional embedding was extracted, while i-vectors, based on GMM-UBM, with a 400-dimension.

A.3.3.B Linguistic modality

We pursued two distinct methods to obtain textual feature embeddings. Firstly, we explored training deep architectures on the relatively small corpus with dimension like the one use in this challenge. Then, we compare this approach with a less data-intensive method based extracting sentence embeddings using a pre-trained model. Both strategies use contextual word embeddings as input but produced different types of learned representations as output. To integrate information from linguistic and acoustic systems, trained architectures were employed to extract linguistic features before the final classification layer, resulting in a single 768-dimensional feature vector for an entire description. In contrast, the sentence embedding approach yielded a 768-dimensional vector for each sentence in a description.

For both approaches, the initial pipeline step involved normalising the data in the ADReSS dataset. Clean transcriptions were encoded into 768-dimensional context embedding vectors using a pre-trained

	Accuracy	Precision	Recall	F1 Score
x-vectors_Vox	0.6818	0.6834	0.6919	0.6812
x-vectors_SRE	**0.7273**	**0.7273**	**0.7273**	**0.7273**
i-vectors_Vox	0.6818	0.7292	0.6818	0.6645
i-vectors_Vox_x-vectors_Vox	0.7273	0.7273	0.7273	0.7273
i-vectors_Vox_x-vectors_SRE	0.7273	0.7351	0.7273	0.7250

Table A.5: Results of different acoustic approaches on the development set

	Accuracy	Precision	Recall	F1 Score
Global Max Pool.	0.7727	0.7947	0.7728	0.7684
LSTM-RNNs	0.8182	0.8182	0.8182	0.8182
LSTM-RNNs Pos	0.8636	0.8667	0.8637	0.8634
GMax/LSTM-RNNs/LSTM-RNNs-Pos	**0.9091**	**0.9091**	**0.9091**	**0.9091**
Sentence emb. - maj. vote	0.7727	0.7947	0.7728	0.7684

Table A.6: Results of different linguistic approaches on the development set

English BERT model with 12 layers and 768 hidden units. The first system, derived from the ComParE2020 Elderly Challenge baseline [318], involved training three neural models on top of contextual word embeddings: (i) a Global Maximum pooling, (ii) a bidirectional LSTM (biLSTM) with an attention module, and (iii) the second model augmented with part-of-speech (POS) embeddings. The loss was evaluated during training on the development set.

The second system, do not require additional training phase, as representations are extracted from a pre-trained model to directly characterise linguistic deficits in AD. Contextual word embeddings obtained for each word were used to compute fixed-size embedding vectors for each sentence by averaging the second to twelfth hidden layers of each word.

A.3.4 Results

The results obtained using the different acoustic feature embeddings are summarised in Table A.5. Different independent models were explored, and an early fusion of the best acoustic results was performed. The x-vectors Voxceleb model generally achieved lower classification accuracy; however, when combining i-vectors and x-vectors extracted from this model, the fused accuracy was comparable to x-vectors trained on the SRE corpus, representing the best result on the development set. These outcomes are slightly lower than those reported in similar works in the literature [297, 313]. However, our approach, distinct from these previous studies as we uses a smaller dataset and do not rely on DNN training. To validate these results on the test set, we select the acoustic feature embeddings extracted from the pre-trained x-vectors SRE model for evaluation.

The results obtained with our various linguistic systems are shown in Table A.6, presenting the performance for features trained with three neural models, their fusion, and the sentence embeddings

	Class	Accuracy	Precision	Recall	F1 Score
Fusion of system	AD	**0.8125**	0.9412	0.6667	0.7805
	non-AD		0.7419	0.9583	0.8364
Sentence embedding	AD	**0.7292**	0.8235	0.5833	0.6829
	non-AD		0.6774	0.8750	0.7636
x-vectors_SRE	AD	**0.5417**	0.5417	0.5417	0.5417
	non-AD		0.5417	0.5417	0.5417

Table A.7: Results of different acoustic and linguistic approaches on the test set

approach. For the sentence embeddings approach, the accuracy use a majority voting strategy over the entire description. Our best classification result attained an accuracy of 90.91% on the development set using the fusion of the linguistic features sets generated by the three neural models. Comparing this result with the one obtained by sentence embeddings, we acknowledge that neural models outperform simpler strategies even with constrained training data. This was somehow surprising and in contradiction with similar experiments performed with the acoustic system. We hypothesise that the large amount of contextual information provided by the Bert model is helpful in overcoming the limited size of the ADReSS dataset. Nevertheless, we suspect that the high accuracy attained with neural models may be too optimistic, due to the fact of having used the development set both for testing and evaluating the model's loss. Thus, in spite of their lower outcome, the sentence embeddings approach is selected as one of the systems to be evaluated on the test set. In fact, on the one hand, we think that they may represent a more reliable system, since do not require additional training. On the other hand, we also observe that they achieve higher classification scores, when compared with a similar approach based on GloVe embeddings [38], thus corroborating our decision.

For a comprehensive evaluation of speech and language impairments in AD, we performed an early fusion of the best results from both the acoustic and linguistic systems. This involved merging x-vectors with linguistic feature sets from three neural models. However, results on the development set using this extended feature set did not yield additional improvements. Despite this, we selected the combined system as our primary choice for evaluation.

For the evaluation, three systems were submitted: (i) a fusion of the best results from linguistic and acoustic systems, (ii) sentence embeddings, and (iii) the best acoustic system. Results on the test set, presented in Table A.7, showed a consistent drop in performance compared to the development set, even for systems not requiring a training phase. The first system achieved the best result with an accuracy of 81.25%, indicating the capability of deep architectures with contextual word embeddings to overcome dataset limitations. The acoustic system alone yielded the lowest accuracy at 54.17%, suggesting room for improvement in adapting acoustic pre-trained models, for example, to better model elderly speech characteristics.

A.4 Transfer Learning-Based Cough Representations for Automatic Detection of COVID-19

A.4.1 Introduction

Finally, we further extend the idea of using pre-trained representation to automatically detect COVID-19 from cough recordings. Indeed, the COVID-19 respiratory disease was declared a pandemic by the World Health Organisation on March 2020, with profound personal, societal, and economic consequences. Clinical diagnosis primarily relies on RT-PCR and antigen tests, but these methods have drawbacks such as significant costs, intrusive sample collections, and delays in diagnosis due to laboratory saturation. To address these challenges, there is a growing interest in developing reliable, cost-effective, immediate, and user-friendly tools to optimise screening campaigns for health care operators, institutions, and companies.

In our work [319], we contributed to the ComParE 2021 COVID-19 Cough Sub-challenge [320]. Firstly, we employ transfer learning to develop COVID-19 classification subsystems using deep cough representation extractors, including TDNN-F and CNN embeddings, as well as PASE+ features. Secondly, we integrate individual decisions from the three experts into a calibrated decision-level fusion system. This ensemble of expert subsystems, relying on cough representations, aims to generate well-calibrated log-likelihood scores across various operating points. The resulting output can be readily interpreted by human experts and seamlessly incorporated into the decision-making process.

Current research on the automatic detection of COVID-19 from speech or respiratory sounds builds on prior studies demonstrating the distinct effects of various respiratory diseases on these sounds. This approach has proven effective in detecting pertussis, asthma, pneumonia, tuberculosis, among others [321]. While conclusive evidence is still pending for COVID-19, preliminary findings suggest specific signatures of COVID-19 in coughs and speech that could potentially enable detection even in apparently asymptomatic individuals and differentiate it from other common respiratory illnesses. Given the limited availability of labeled COVID-19 data, many approaches rely on transfer learning, data augmentation, and class balancing techniques.

The majority of previous work in this domain relies on CNNs. For instance, a pre-trained VGGish model [322]is employed as a generic audio feature extractor in [323]. Other works fine-tune CNNs initially trained for cough detection for the purpose of COVID-19 detection [324, 325]. Ensemble models incorporating both DNNs and CNNs, directly trained from scratch for COVID-19 detection, have also been proposed [326]. Additionally, certain studies [327] leverage information about self-reported symptoms, encoding them as one-hot vectors and combining them either at the feature- or decision-levels with traditional speech features.

A.4.2 Corpora

In this work, we used two datasets: the COVID-19 COUGH (C19C) corpus, provided in the ComParE 2021 COVID-19 Cough Sub-Challenge [323, 327] for evaluation, and the COUGHVID corpus [328], employed for training and fine-tuning transfer learning-based cough representation extractors. Silence segments were eliminated from both datasets.

The COVID-19 COUGH (C19C) corpus is a subset of the Cambridge COVID-19 Sound database [323, 327], consisting of 725 cough recordings from 397 participants with self-reported COVID-19 status labels (positive/negative). The corpus is distributed into train (71 positives/215 negatives), development (48 positives/183 negatives), and a blind test set (208 samples) with gender-balanced subsets.

In our preliminary analysis, it was observed that some files had a reduced bandwidth of 4 kHz, potentially corresponding to samples originally recorded at 8 kHz. Namely, 13, 8 and 8 narrow-band files were detected in the train, development and test subsets, respectively. This condition certainly reflects the reality of many real-world applications. However, we noticed that all the narrow-band recordings in the train and development subsets correspond to the COVID-19 positive class To address this, a second version of the dataset, denoted as "$C19C_{fullband}$", was created by removing narrow-band recordings from the original train and development subsets. This resulted in 273 samples in the train subset (58 positives/215 negatives) and 223 in the development subset (40 positives/183 negatives), while the test subset remained unchanged for consistent challenge evaluation conditions.

The COUGHVID corpus [328] is a publicly open dataset consisting of non-curated recordings performed using lossy codification. The dataset includes a variety of conditions such as sampling rate, bandwidth, number of channels, and quality. Volunteers recorded their coughs and reported their COVID-19 status (positive/symptomatic/healthy), age, gender, and medical condition. The dataset comprises 27,550 recordings, with 15,125 classified as coughs by an automatic cough detector. Of these, 10,763 have self-provided gender and COVID-19 status annotations, including 680 COVID-19 positives (395 male/285 female), 8,270 healthy (5,632 male/2,638 female), and 1,813 symptomatic (1,114 male/699 female). Additionally, a small fraction of the dataset was annotated by expert pulmonologists with information on various aspects, such as type of cough, presence of audible symptoms, diagnosis, and severity.

A.4.3 Proposed system

A.4.3.A TDNN-F embedddings

X-vector embeddings are currently regarded as state-of-the-art speaker representations, surpassing other proposed representations like d-vectors and i-vectors. Motivated by the results of the two previous works mentioned in this annex, this study explores the applicability of x-vector-like embeddings to coughs, aiming to encode relevant information about the cough signal for medical insights. The X-vector extractor is implemented using TDNN-F as proposed for speaker recognition in [329]. Cough embeddings are 128-dimensional vectors obtained at the output of the final dense layer. The network undergoes two-stage training: initially an age estimation and gender classification using a subset of the COUGHVID dataset was performed, and subsequently fine-tuning with expert-annotated data for tasks closely related to COVID-19 classification. These tasks include cough type, presence of dyspnea, presence of wheezing, diagnosis, and severity. The reason behind this fine-tuning step is the fact that these tasks are much closer to COVID-19 classification than age and gender. Input features consist of 30 MFCCs computed every 10 ms from 25 ms-length frames, following the egs/voxceleb/v2 Kaldi recipe [69].

A.4.3.B CNN embedddings

CNN-based approaches for COVID-19 detection suffer from some limitations such as the use of CNNs as generic audio feature extractors without task-specific tuning or relying on relatively small datasets for training. In contrast, our work addresses these limitations by leveraging transfer knowledge from the VGGish model, originally trained on a vast corpus for audio classification, and subsequently fine-tuning it for COVID-19 detection using the COUGHVID dataset.

The VGGish model [322], is an adaptation of the VGG network [330] for audio classification. It consists of four blocks with convolutional and pooling layers, followed by fully-connected layers and an output layer. In this work, a simplified version of the model is used and was pre-trained using 5.4 million hours from YouTube data. For our experiments, we used two different settings. In the first setting, the model serves as a pre-trained generic feature extractor with weights directly loaded from the original model. In the second setting, the model is fine-tuned for COVID-19 detection using a balanced subset of the COUGHVID dataset. The input to the VGGish network is log Mel-spectrogram features computed every 0.24 s from 0.96 s-length segments amd the resulting embeddings are 256-dimensional vectors.

When fine-tuning, layers 9 and 10 are included to facilitate training with limited data, with their weights initialized randomly. The entire CNN is fine-tuned for 150 epochs using cross-entropy loss, the Adam optimizer with a learning rate of 10^{-5}, and a batch size of 64. The fine-tuning is conducted on a balanced subset of the COUGHVID dataset, consisting of 680 positive and 680 negative cough recordings, with 80% used for training and 20% for development.

A.4.3.C PASE+ embedddings

The study incorporates the problem-agnostic speech encoder model, PASE+ [331,332], g where targets are learned directly from the signal. PASE+ features are derived from a shared encoder with a SincNet-based layer [90], seven convolutional blocks, and a Quasi-RNN layer. The encoder output is connected to twelve workers, each designed for specific tasks like reconstruction of waveform, Log Power Spectrum (LPC), MFCCs, prosody, fbanks, gammatone, and binary discrimination tasks. Two PASE+ extractors were

System	*dev*	$dev_{fullband}$	*test*
ComParE 2021 CCS Sub-challenge Baseline			
openSMILE	61.4	53.0	65.5
openXBOW$_{2000}$	64.7	56.5	72.9
DeepSpectrum+SVM	63.3	57.3	64.1
auDEEP$_{\text{-60 dB}}$	67.6	57.3	67.6
End2You	61.8	-	64.7
Fusion of Best	-	-	73.9
TDNN-F Embeddings			
Trained COUGHVID$_{\text{Step1}}$	68.8	63.6	-
Fine-tuned COUGHVID$_{\text{Step2}}$	68.1	62.3	-
CNN Embeddings			
Pre-trained YouTube	66.9	62.4	-
Fine-tuned COUGHVID	71.2^{+}	65.6	62.3^{+}
PASE+ Features			
Trained Librispeech	63.1	61.7	-
Trained COUGHVID	67.4	**66.8**$^{+}$	64.1^{+}
Calibrated Fusion			
Fusion of experts	**72.3**$^{+}$	66.1	69.3^{+}

Table A.8: Performance results (unweighted average recall-UAR) on the COVID-19 COUGH (C19C) corpus

used: one pre-trained on Librispeech and another trained from scratch on COUGHVID data. Both are trained for 150 epochs, creating 256-dimensional feature vectors for each frame every 10 ms.

A.4.3.D COVID-19 condition classification

The study employs transfer learning to address the limited COVID-19 data for training. Three SVMs we used on TDNN-F embeddings, CNN embeddings, and PASE+ features, respectively, for expert decisions. TDNN-F embeddings are directly input to the SVM. CNN-based embeddings are derived from cough segments and subjected to majority voting for the final decision. PASE+ features are averaged across the sequence and fed to the SVM classifier. The SVMs are trained on both the C19C and C19C$_{fullband}$ datasets, exploring various kernels, data normalisations, and class balancing methods. Hyperparameters are optimised through grid-search on development subsets, and linear logistic regression is applied to combine system decisions with scaling factors. The regression approximates log-likelihood ratios, thus, a theoretically determined decision threshold can be used for making hard decisions.

A.4.4 Results

Table A.8 shows the comparison between ComParE 2021 CCS baselines and our proposed system. All systems were separately trained on both the C19C and C19C$_{fullband}$ subsets, and evaluations were conducted on the corresponding dev and dev$_{fullband}$ subsets. The reported test results are based on the best individual systems trained on the C19C and C19C$_{fullband}$ datasets (marked with +), presented in terms of Unweighted Average Recall (UAR).

The proposal demonstrates competitive performance compared to baseline systems. The TDNN-F x-vector embeddings-based expert, trained initially on gender classification and age regression tasks using COUGHVID data, achieves a UAR of 63.6% on devfullband. However, fine-tuning in step 2 using a multi-task setting does not significantly enhance cough representations, suggesting potential overfitting for some subtasks with limited data.

The CNN embeddings pre-trained on YouTube videos exhibit reasonable performance, improving to 65.6% UAR after fine-tuning with COVID-19-specific data. The PASE+ features, trained with COUGHVID data, yield the best performance among the three expert systems, with a development UAR of 66.8% and a test UAR of 64.1%. Notably, the PASE+ extractor trained on the larger Librispeech dataset achieves a 5.1% absolute improvement inUAR when trained with COVID-19-specific data, indicating the significant benefit of such data.

The fusion of the best x-vector (Trained COUGHVIDStep1), CNN (Fine-tuned COUGHVID), and PASE+ (Trained COUGHVID) experts yields a UAR of 72.3% on development (1.1% absolute improvement from the best expert) and 69.3% on test when trained on the C19C datasets. The underperformance on the C19C$_{fullband}$ subset warrants further analysis but may be attributed to the low number of COVID-19 positive examples.

A.5 Conclusion and future work

In the three distinct research works done within the context of this thesis, we employed pre-trained embedding extractors as tools for detecting pathologies. These embeddings can be used in two different way: functioning either directly as feature extractors or undergoing fine-tuning for specific tasks. Our findings in these works present promising results, indicating that embeddings trained on substantial amounts of data may contain valuable health-related information.

We began by exploring the replacement of knowledge-based features with task-agnostic speaker representations in multiple disease detection. Focusing on x-vector embeddings trained with elderly speech data, our experiments with European Portuguese datasets supported the hypothesis that discriminative speaker embeddings, particularly x-vectors, contain relevant information for disease detection that knowledge-based features may fail to represent. Notably, x-vectors proved more suitable than i-vectors for tasks with domain mismatches, such as verbal task mismatches and cross-lingual experiments. Future endeavors include training the x-vector network with augmented and multilingual datasets, extending the approach to other diseases and verbal tasks, and replicating experiments with in-the-wild data collected from online multimedia repositories.

Moving to the context of AD classification, we adopted a multi-modal approach, leveraging automatically learned feature representations. Our investigation covered both acoustic and linguistic, exploring feature embedding vectors from pre-trained models and training deep neural architectures. By combining these approaches, we achieved an accuracy of 90.91% and 81.25% on the development and test sets, respectively. Notably, acoustic systems demonstrated a greater need for data to improve predictive ability, especially in the presence of potential ASR errors. Future work may involve analysing the impact of ASR errors, exploring robust acoustic methods tailored to AD speech characteristics, and delving into the implications of atypical speech.

Lastly, our focus shifted to the ComParE 2021 COVID-19 Cough Sub-challenge. Leveraging transfer learning, we developed three expert classifiers: TDNN-F embeddings, CNN embeddings, and PASE+ features. While our results demonstrated competitive performance compared to baseline systems, caution is warranted due to the limited data. To enhance the reliability of COVID-19 screening tools, larger datasets are recommended for better learning of cough representations, coupled with more suitable back-end classifiers. Future exploration could includes recurrent neural networks with attention mechanisms to capture temporal dynamics, SSL, multi-task TDNN-F network-based cough embeddings, and assessing the suitability of PASE+ features as an alternative input representation.

Collectively, these contributions showcase the diverse applications of advanced technologies in addressing challenges in health-related tasks using pre-trained representations, ranging from COVID-19 screening through cough analysis to AD classification and beyond, ultimately paving the way for impactful advancements in SLT.

B

Sefl-supervised learning as feature extractor for children's ASR

B.1 Introduction

In light of the increasing availability of large amount of unlabeled speech data, there has been a need to efficiently extract general-purpose knowledge from it. Consequently, SSL has experience notable advances and has gained substantial attention in recent years especially in the context of low-resource tasks. SSL refers to a training paradigm where a model learns representations from unlabeled speech data without relying on explicit labels or annotations. This approach yields discernible enhancements in performance across various applications such as speech recognition, speaker identification, and emotion recognition [163]. Traditionally, SSL models can be used in two manners: firstly, as a feature extraction to replace human-designed features [333, 334], and secondly, as a model initialisation accomplished by concatenating prediction layers and fine-tuning the entire model [167, 169, 171, 236].

In recent years, an incrase number of research focus on using SSL models for children ASR. One notable avenue observed in the literature involves the fine-tuning of SSL models exclusively using children's data [170] or through a combination of adult and children's data [169]. These efforts have yielded improved performances, showcasing the adaptability of SSL models to the nuances inherent in the speech patterns of children. Conversely, an alternative strategy has emerged by using Adapters to adapt SSL models specifically to children's speech characteristics subsequently followed by the full model, Adapters included, fine-tuning [171]. This initial training, allow to reduce the complexity associated with the conventional process of fine-tuning the entire model directly with children speech. The observed successes of SSL as an initialisation for fine-tuning in the context of children's ASR underscore its efficacy as a new training paradigm. However, the exploration of SSL models as front-end feature extractors, compared to the conventional hand-crafted features, despite its use and success for adult speech [333, 334] remains unexplored for childen's ASR. Motivated by these considerations, this chapter undertakes a comprehensive review of various SSL models, evaluating their potential as novel feature extractors in the domain of children's ASR.

B.2 Self-supervised pre-trained models

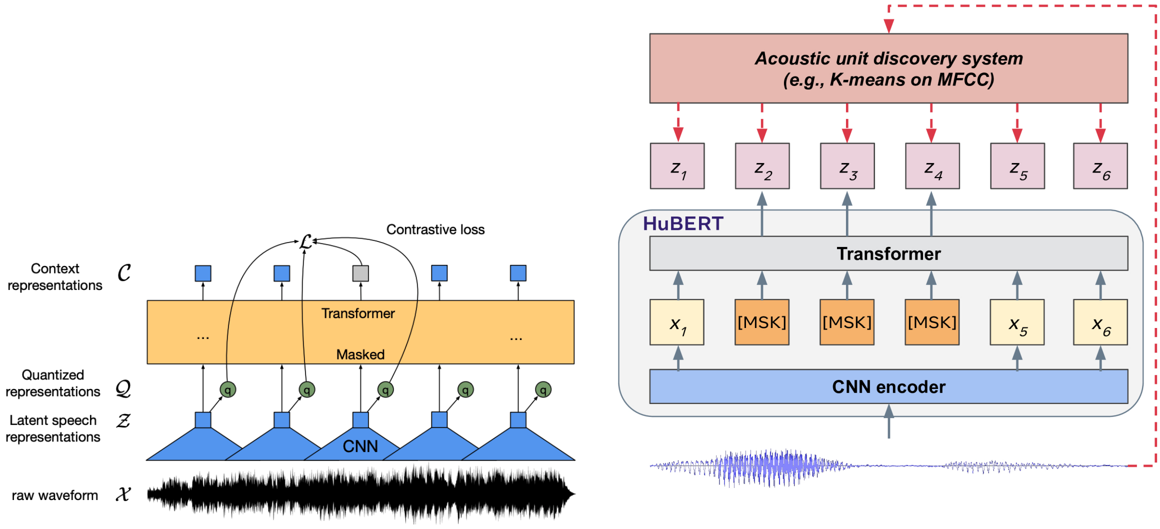
Method	Architecture	#Params	Stride	Input	Corpus
FBANK	-	0	10ms	waveform	-
Mockingjay [335]	12-Trans	85.12M	10ms	FBANK	LS 360 hr
Audio ALBERT [336]	3-Trans	7.15M	10ms	FBANK	LS 960 hr
NPC [337]	4-Conv, 4-Masked Conv	19.38M	10ms	FBANK	LS 360 hr
APC [338]	3-GRU	4.11M	10ms	FBANK	LS 360 hr
TERA [339]	3-Trans	21.33M	10ms	FBANK	LS 960 hr
wav2vec 2.0 Base [163]	7-Conv, 12-Trans	95.04M	20ms	waveform	LS 960 hr
wav2vec 2.0 Large [163]	7-Conv, 24-Trans	317.38M	20ms	waveform	LL 60k hr
Distill HuBERT [239]	7-Conv, 3-Trans	23.49M	20ms	waveform	LS 960 hr
HuBERT Base [164]	7-Conv, 12-Trans	94.68M	20ms	waveform	LS 960 hr
HuBERT Large [164]	7-Conv, 24-Trans	316.61M	20ms	waveform	LL 60k hr

Table B.1: Overview of different SSL architectures used as frozen feature extractors

Traditionally, SSL models can be categorised into two distinct approaches: generative modeling and discriminative modeling. In this section, we will focus on summarising a selection of SSL models, presented in Table B.1, with a particular emphasis on their differences.

B.2.1 Generative modeling

Generative modeling has emerged as a prevalent approach for learning speech representations for SSL. Generally, generative models are trained to generate speech frames based on their learned speech representations with or without the help of context. For instance, the Autoregressive Predictive Coding (APC)



(a) Illustration of the Wav2vec2 architecture taken from [163]

(b) Illustration of the HuBERT architecture taken from [164]

Figure B.1: Overview of the discriminative SSL Wav2vec2 and HuBERT models

model [338] adopts a language model-like training paradigm, where a RNN generates future frames predicted from the past frames. The Mockingjay model [335] takes inspiration from BERT-like pretraining techniques by masking input acoustic features along the time axis and subsequently regenerating the masked frames while the Audio ALBERT represent a smaller version of it [336]. Expanding upon this concept, the Temporal Encoder Representations from Acoustics (TERA) model [339] introduces an additional layer of complexity by masking bins in the frequency axis alongside the temporal axis. Finally, the Non-autoregressive Predictive Coding (NPC) model [337] combines elements from both APC and Mockingjay by substituting the RNN in APC with a CNN layers and modifying the future frames generation process into a masked reconstruction.

B.2.2 Discriminative modeling

The success of discriminative modeling has been notably pronounced with the introduction of the contrastive loss, a technique where the model discerns between correlated positive samples and negative samples. The underlying intuition is that positive samples should exhibit closer representations in comparison to their negative counterparts. A prominent exemplar of this approach is evident in the Wav2Vec2 model [163], which has demonstrated promising potential in learning speech representations. The Wav2Vec2 model achieves this by masking latent representations of the raw waveform and formulating a contrastive task over quantized speech representations. A more detailed representation of Wav2Vec2 architecture is displayed in figure B.1(a). Later, moving away from the contrastive loss, the work of [164] proposed the HuBERT model. This model introduces a novel methodology by incorporating BERT's token prediction via offline clustering on representations. Specifically, the HuBERT model use a BERT-like training that consumes masked continuous speech features to predict pre-determined cluster assignments. The labels assigned to the masked locations during clustering serve as the predicted targets. Importantly, the predictive loss is selectively applied solely over the masked regions, compelling the model to learn robust high-level representations of unmasked inputs in order to accurately infer the targets of the masked ones. The HuBERT architecture is shown in figure B.1(b). An distilled version of the HuBERT is also explored in our experiments [239].

B.3 Experimental setup

In our experimental setup, we used the Self-Supervised Speech Pre-training and Representation Learning (s3prl) toolkit[1]. The s3prl allow the modular use of pre-trained SSL models, called upstream, to perform various downstream tasks. In order to evaluate the efficiency of different SSL models as features extractor for children's ASR, we froze the pre-trained models's weight to extract embedding representation of the speech signal as new acoustic features. All the differents SSL upstream models used in our experiments are listed along with detailed informations regarding their architectures in table B.1. Notably, each of these models underwent self-supervised pre-training on either 360 or 960 hours of LibriSpeech [31] (denoted as LS 360hr and LS 960hr, respectively) or on an extensive 60 thousand hours of LibriLight data [340] (referred to as LL 60k hr) . The ASR downstream task was conducted using a 2-layered Bidirectional LSTM (BiLSTM) architecture with 1024 units, optimised with a CTC loss. The training spanned 800 thousand iterations, with a learning rate set at $1.0\dot{1}0^{-4}$. Additionally, a dropout rate of 0.2 was applied to the BiLSTM architecture to increase robustness. Limited by the large size of some SSL model, we decided to used a subset of the Myst [33] dataset, using 77 hours of speech for training, by removing the longest utterances in the train and validation sets. A detailed description of the filtered Myst data is provided in table B.2.

	Training	Validation	Test
# of utterances	23594	3959	4079
# of speakers	559	79	91
# of hours	77	12	13

Table B.2: My Science Tutor Children Speech Subset Corpus statistics

Model type	SSL upsteam	UER↓	WER ↓
Hand-crafted	Fbanks	12.29%	35.14
Generative	Mockingjay	12.49%	35.08%
	Audio Albert	12.28%	34.69%
	NPC	11.99%	33.07%
	APC	11.88%	32.84%
	TERA	11.31%	31.80%
Discriminative	Wav2Vec2 Base	7.37%	19.76%
	Wav2Vec2 Large	7.00%	18.76%
	Distill HuBERT	9.22%	25.75%
	HuBERT Base	7.40%	19.77%
	HuBERT Large	**6.03%**	**15.41%**

Table B.3: Results without language model of different Self-supervised models as feature extractors

B.4 Results

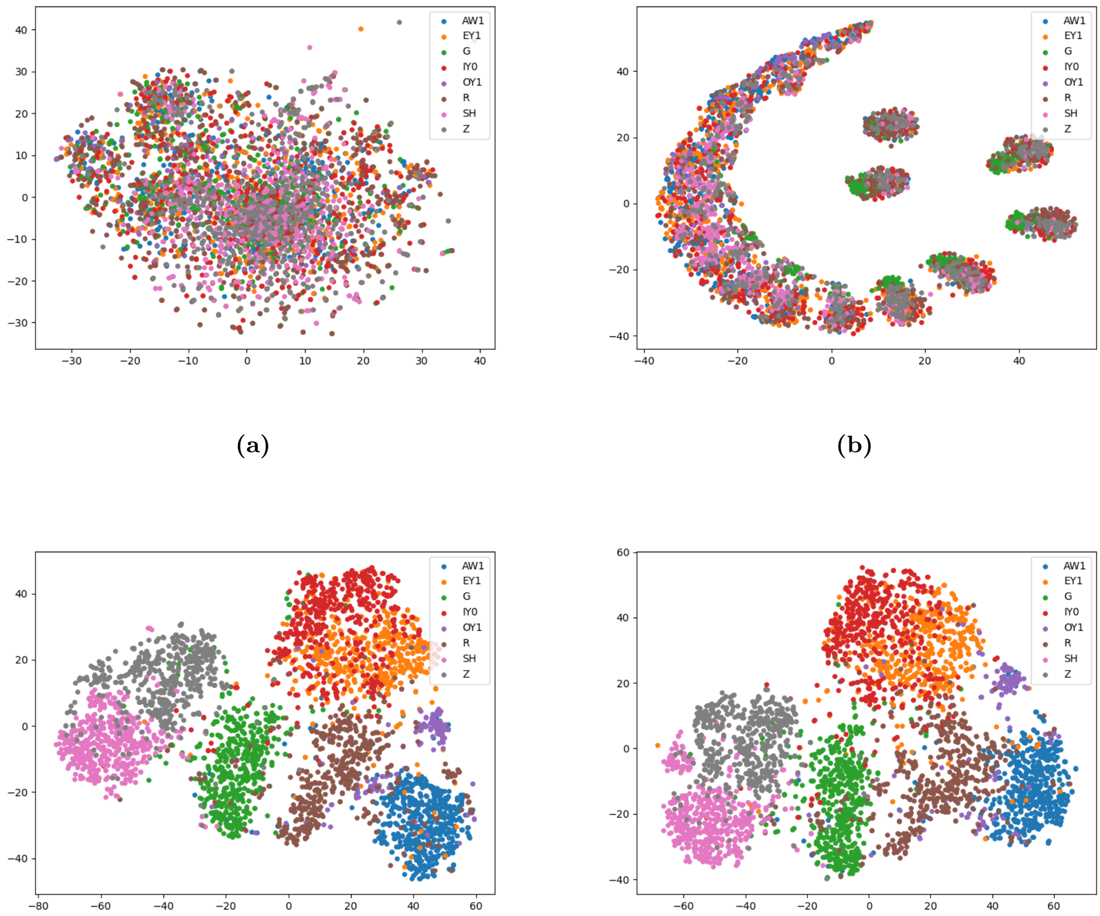
Table B.3 present the results of the comparaison between various SSL pre-trained model as feature extractors for children's ASR. We provide Unit error rate (UER) as well as WER. fbanks are established as a baseline, yielding a UER of 12.29% and a WER of 35.14%. In terms of generative SSL models, TERA and Audio Albert surpass traditional fbanks, exhibiting improvements in both UER ,11.31% and 12.28% respectively and WER with 31.80% and 34.69% respectively. Turning to discriminative SSL, the Wav2Vec2.0 Base and Wav2Vec2 Large demonstrate substantial enhancements in performance, achieving UER values of 7.37% and 7.00%, and WER of 19.76% and 18.76%, respectively. The distilled version of HuBERT outperforms fbanks but falls behind the Wav2Vec2 models in terms of both UER and WER with 9.22% UER and 25.75% WER. Finally, HuBERT Base and HuBERT Large emerge as the top-performing SSL models, boasting the lowest UER with respectively 7.40% and 6.03% and WER of 19.77% and 15.41%. We observed that the best performing models are the large discriminative pre-trained on a large amount of speech data.

The results suggest that large Discriminative models, particularly HuBERT Large, demonstrate superior performance compared to other SSL models and fbanks as featuress extractor for children ASR. Notably, even though no language model where used in this experiment, the results are of the same order as those reported in previous the different chapters of the thesis obtained with a transformer and a transformer language model. Showing the benefic of using frozen pre-trained SSL models as feature extractor for children ASR.

B.5 Analysis of the extracted features

In this section, we delve in more detail into the distinct features extracted from various models, aiming to better understand the notable performance differences observed between Wav2Vec2 and HuBERT,

[1] https://github.com/s3prl/s3prl



(a) **(b)**

(c) **(d)**

Figure B.2: (a) Fbanks (b) TERA (c) Wav2Vec2 (d) HuBERT
T-SNE plot of the different extracted features using the same speech data using phoneme labels

in contrast to generative models like TERA and traditional filterbanks. In a first step we aligned our children's speech data to obtain phoneme alignments using the Montreal Forced Aligner [341].

Therefore, for each phonemes present in a utterance we obtained a variable-length feature sequences corresponding the extracted features for the different frames where the phoneme has been aligned. In order to get a single vector for each phoneme, we average these sequences. It is noteworthy that silence frames have been excluded. This operation is repeated for a subset of one thousand utterances of children speech in order to gather different exemple of the same phoneme from different speaker and different context. Subsequently, t-Distributed Stochastic Neighbor Embedding (t-SNE) plots are generated for each of the studied models.

The t-SNE plots are depicted in Figure B.2. We observe that traditional filterbanks features form a cloud points witn no structure. This lack of structure suggests that the fbanks features do not inherently exhibit phoneme-related information. Moving on to the t-SNE plot for TERA features, clusters are observable, but these clusters do not align with phoneme. This observation indicates that while TERA features capture information from speech, they not encode phoneme-specific information, as evidenced by the presence of mixed phonemes within the different clusters. In contrast, both Wav2Vec2 and HuBERT exhibit highly similar plots, wherein distinct clusters corresponding to different phonemes are evident. This finding suggests that the features extracted from Wav2Vec2 and HuBERT inherently capture phonemic information, even though no explicit phoneme annotations were provided during training. The presence of these phoneme-related clusters indicates that the usage of these features facilitate the ASR task by implicitly encoding phoneme information.

B.6 Conclusions and future work

However, it is crucial to emphasise that the outcomes and findings of our experiments are highly dependent on the language used. Specifically, all the SSL models were pre-trained using English language, which align with the language of the children dataset used in this experiment. Consequently, the generalisability of our results may be limited when applied to different languages. Using a different language could potentially result in decreased performances, as the SSL models may not be as adept at capturing the linguistic nuances and acoustic characteristics specific of that particular language [342]. To address the language-dependent nature of our results, future work could explore the efficacy of employing multi-lingual SSL models, such as XLS-R [214].