TÉCNICO LISBOA

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

Expected 2022

Abstract

Speech is a fundamental communication skill used for social interaction, express feelings and needs, among others. Unfortunately, there is an increasing number of people who suffer from debilitating speech pathologies, including children. A childhood speech disorder that is not properly diagnosed and treated can have long-term negative effects in social, communication and educational situations. Hence the importance of speech therapy.

In recent years, machine learning has proven to have many applications in the health field, both for diagnosis and monitoring. Particularly, there is a growing interest in the development of automatic tools to assist speech-language pathologists. Furthermore, some automatic tools allow patients to do exercises outside the session with the therapist, for example at home. In the case of children, the exercises can be provided in a gamification context, which contributes to motivate them to practice more often. Most of these tools require reliable automatic speech recognition systems, as these are commonly used to provide pronunciation quality scores. However, despite recent significant improvements in the performance of automatic speech recognition systems for healthy adults, these systems' performance dramatically drops when considering speech from children and speakers with speech pathologies. This is mainly due to the high acoustic variability and the reduced amount of training data available.

In this proposal, we propose to investigate how machine learning algorithms in speech recognition tools for children can enable more complex exercises and better feedback. We first present a summary of existing automatic speech recognition techniques for children. In view of the need for a robust oracle model for children, this proposal will mainly focus on improving the automatic speech recognition system for children. Therefore, we identified several methods for improvement. i) Knowledge transfer on Hybrid and end-to-end models by using transfer and multi-task learning and ii) Adapter transfer in end-to-end speech recognition for a parameter efficient transfer.

Keywords

Automatic Speech Recognition Children, Children speech, Atypical speech, Atypical speech recognition

Contents

List of Figures

List of Tables

1

Introduction

Contents

1.1 Context

The faculty of expressing or describing thoughts, feelings and needs by using language is a fundamental ability 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 a 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 conceptualization of the message to be transmitted (*conceptualisation*), followed by the selection of appropriate lexical items and subsequent grammatical encoding, culminating by 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*) [11]. Unlike language, which encompasses both spoken and written forms, speech specifically refers to the spoken manifestation of language.

This 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 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 [12] highlighted that 7.7% of children aged 3 to 17 in the United States exhibited communication disorders, with 5.0% of this cohort specifically presenting speech-related problems.

Furthermore, findings [13] suggests that individuals afflicted 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 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 tailor 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, the setting 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 [14, 15]. Teletherapy, for instance, has emerged as a promising avenue to bridge geographical gaps and deliver therapy services remotely [16].

In this context, Speech and Language Technologies (SLT) have emerged as highly pertinent within the domain of speech therapy [17]. 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 [18]. 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 SLT examples

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.

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

1.2 Problem statement

Recent years have seen an 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 performances progress in these applications was made possible through the use of machine learning techniques, espacially 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, limiting their capacity to fully benefit from digital platforms. Children commonly face difficulties to manipulate 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.

As previously mentioned, SLTs 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 cutting-edge systems with a focus on assisting SLPs. In this context, systems capable of automatically recognising speech

content, assessing pronunciation quality and detecting speech pathologies could be highly valuable in supporting pediatric SLPs and patients.

All of these objectives require the implementation of a robust automatic speech recognition (ASR) system specifically tailored for healthy children, serving as a foundational model. Nevertheless, while speech recognition technologies have made substantial advancements, leading to increased accuracy, the performance of ASR systems for children remains underperforming in comparison to their adult-oriented counterparts. This discrepancy results leads into unreliable systems for children's speech. The diminished performance can be atributed 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 methodologies for enhancing ASR systems specifically designed to accommodate the variability present in children's speech. The overarching aim is to establish a robust foundational system that effectively addresses the recognition of children's speech characteristics.

Our work specificialy 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 an speaker-based, parameter-efficient automatic speech recognition model?

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 began with a thorough exploration of the current state-of-the-art of children's ASR. The primary objective was to identify the various avenues by which improvements could be envisaged throughout this thesis. The state-of-the-art review constituted a thorough examination of existing literature, research papers, and technological advancements related to children's speech processing in general. The aim was

to understand the fundamental determinants that contribute to the decline in ASR performance for children's speech. By meticulously assessing current research on children speech, we identified challenges and potential areas for possible impact.

Subsequent to the exhaustive literature review, our research transitioned into 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 childrren and adult 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 adapt speech recognition adult models, fine-tuning them for children's speech. In the other hand, multi-task learning exposed models to both adult and children's speech datasets simultanously during training. In an innovative synthesis, we combined 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:

- **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 to the end-to-end paradigm, motivated by the encouraging improvements observed in the end-to-end children's ASR performance. By adopting a detailed transfer learning approach, we aimed to gain a comprehensive understanding of the specific components of the end-to-end architecture that proved most relevant and played a central role in these notable score enhancements. The identification of the most relevant components allowed the development of specific algorithms aimed at further improving the model. Particularly, we explored the integration of an additional set of parameters directly into the original ASR model. This integration facilitated a parameter-efficient approach to fine-tuning the model:

- **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 REVIEW.

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:

- **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 REVIEW.

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

- Pompili, Anna, **Thomas Rolland**, and Alberto Abad. "The INESC-ID multi-modal system for the ADReSS 2020 challenge." *Interspeech* (2020).
- Botelho, Catarina, Francisco Teixeira, **Thomas Rolland**, Alberto Abad, and Isabel Trancoso. "Pathological speech detection using x-vector embeddings." *arXiv preprint* arXiv:2003.00864 (2020).
- Solera-Ureña, Rubén, 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 five chapters. In Chapter 2, a comprehensive review of related work is conducted to establish the context and understanding of the challenges associated with automatic children's speech recognition. Furthermore, we provides an overview of the history of automatic speech recognition systems, along with an examination of the latest approaches specifically tailored to address the unique challenges posed by children's ASR. Finally, a compilation of children's speech corpora, available online and referenced in prior literature, is presented.

Following this, Chapter 3 we present our work done on the hybrid speech recognition framework.

2

Background - Children automatic speech recognition

Contents

Automatic Speech Recognition (ASR), or Speech-to-text (STT) refers to the process of mapping a raw spoken audio utterance into its corresponding text. The potential use case of ASR in nu,erous application across various domain drived the demand for robus and reliable ASR systems. ASR applications extend across diverse sectors including academic, medical, industrial or military fields. The momentum in ASR innovation has made significant progress in the recent years, thanks to the attention and investment from both industry and government sectors. This support has led to the release of applications such as voice assistants, hands-free interfaces, healthcare assistance, live translation, and more.

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. The focus on adult speech is explained by the potential immediate applications of ASR systems for this target audience. In addition, training ASR models on adult speech has advantages in terms of data availability and the relatively stable aspects of adult speech characteristics. Indeed, adult speech is often more standardised, with established linguistic conventions and stable features. As a result, ASR systems trained on adult speech tend to perform better. However, the challenge arises when such systems are applied to recognise speech in mismatched scenarios, like children's speech. For example, in the context of children speech, ASR algorithms exhibit a stark performance decline, often two to five times worse [19]. This discrepancy in performance can be attributed to the intra- and inter-speaker variability. In fact, speech serves as a conduit not solely for linguistic content but also for paralinguistic cues that unveil aspects of the speaker's identity, including age, gender, health status, emotional state, and regional origin. While this additional layer of information is incredibly valuable for human communication, it does introduce complexity and challenges for accurate ASR systems [20]. Moreover, several external factors further negatively impact the performance, including noise, speaker variability, mispronunciation, and the quality of the recording [21, 22].

More recently, the potential applications of automatic speech recognition in education and entertainment have led to a growing interest in automatic speech recognition for children. Indeed, children represent a demographic that aligns well with these applications for several compelling reasons. The complexity of traditional computer interfaces, like keyboards and mice, can pose challenges for young children, making voice interfaces a more accessible and user-friendly option. Furthermore, speech and language applications, including reading tutors and speech and language acquisition assistants, hold the promise of addressing educational inequalities among children and facilitating their integration into society.

In this chapter, we first present the diverse challenges associated with children's ASR. These challenges encapsulate the unique characteristics of children's speech, including high acoustic and linguistic variability, as well as the limited labeled data for training. Following this, we provide a brief introduction to ASR, tracing its historical evolution from early pattern recognition approaches to the advent of statistical models and the contemporary shift towards end-to-end models. This historical context sets

the stage for understanding the underlying principles and advancements in ASR technologies. Next, the chapter transitions to a comprehensive review of state-of-the-art methods specifically designed to address the challenges posed by children's ASR. These methods are examined from diverse perspectives, encompassing acoustic adaptation, as well as advancements in machine learning. This thorough exploration aims to provide a clear overview of the different techniques used to improve children's ASR. Finally, the chapter concludes with a discussion that synthesises the various perspectives presented earlier. Here, we outle the rationale behind the selected methodologies used in the context of this thesis.

2.1 Children speech recognition challenges

In this section, we explore the distinct challenges posed by children's speech to automatic speech recognition (ASR) systems. The divergence between child and adult speech is primarily caused by the ongoing growth and intellectual development of children. In this context, we present three primary factors and hurdles linked to children's ASR. 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 pose a significant challenge for speech recognition systems, which are often trained on adult speech datasets. Accounting for this acoustic variability becomes imperative for the development of accurate and robust ASR models adapted to the unique characteristics of children's speech. Next, we will delve into the complex aspects of linguistic and phonetic knowledge inherent for children. Children undergo dynamic changes in language, with vocabulary expansion, and phonetic development as they grow older. This linguistic evolution poses challenges related to the recognition of age-specific linguistic patterns and variations in pronunciation. Similar to the acoustic variability, effective modeling of these linguistic subtleties is important for children speech recognition systems. Finally, we present the challenge of the limited availability of corpora of children's speech. Unlike adult speech, data corpus containing labeled examples of children's speech are relatively rare and small. This scarcity hinders the training of ASR models, limiting their exposure to the various linguistic and acoustic variations of children's speech.

2.1.1 Speech variability

Speech production is a complex process that involves the synchronized actions of multiple components within the speech production apparatus. These components include the vocal folds, tongue, lips, and mouth,all working collaboratively to generate speech. The coordination of these elements results in fluctuations in air pressure, ultimately producing the speech waveform, which is essentially a measurement of air pressure over time. Human speech waveforms encompass a range of frequency components spanning from 20 Hertz (Hz) to 20kHz, which are detected and processed by the human auditory system. A

precise comprehension of these frequency components, such as the fundamental and formant frequencies, is essential for the development of effective speech-processing tools.

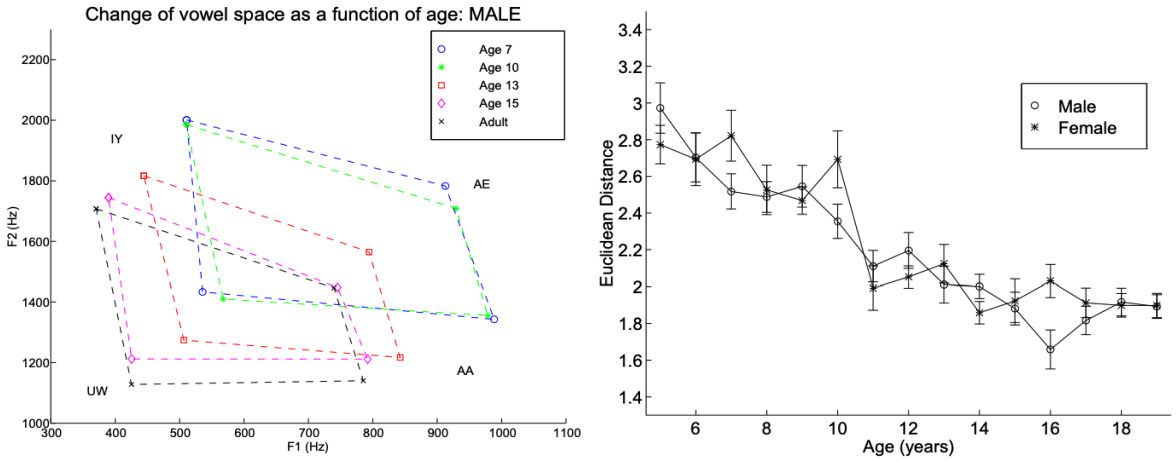
The fundamental frequency, often referred to as F0, holds a crucial role in the analysis of speech signals. It characterises the (quasi-)periodic average oscillations produced by the vibrations of the vocal folds. Measured in Hz, F0 is often referred to as the acoustic correlate of pitch. F0 exhibits an inverse relationship with the vibrating mass of the vocal folds, leading to distinctive F0 values across different demographic groups. Typically, adult men exhibit lower F0 values, ranging from approximately 100 to 150Hz. In contrast, women tend to have higher F0 values, typically falling within the range of 200 to 300Hz. Children, with their smaller vocal folds, often demonstrate even higher F0 values, generally ranging from 300 to 450Hz. These variations in F0 contribute to the perceptual differences in pitch between individuals of different ages and genders. According to [2], significant F0 differences between male and female speakers emerge starting from age 12. For male speakers, the drops in F0 from age 11 to age 15, with no significant pitch change after age 15. This suggests that, on average, pubertal pitch change in male speakers commences between age 12 and 13 and concludes around age 15. The relatively large between-subject variation at ages 13 and 14 also implies that the onset time of puberty varies among speakers in these age groups. For female speakers, the pitch drop from age 7 to 12, and there is no significant pitch change after that age. In addition, the F0 change for female subjects is more gradual compared to male speakers.

It is crucial to highlight that F0 is not a static parameter; instead, it exhibits continuous variation within a sentence. This dynamic nature allows F0 to be used expressively in speech, conveying nuances such as emphasis, emotion, and intonation patterns. The variations in F0 contribute to distinguishing between different types of speech acts, such as declarative statements, questions, and exclamations.

A formant refers to a concentration of acoustic energy centered around a specific frequency within a speech waveform. As per the definition 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 characterizing vowel sounds and are instrumental in distinguishing between different vowels. In speech analysis, the first three formants, denoted as F1, F2, and F3, are commonly utilized for their significance 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 [23] 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 vowel frequencies in children's speech 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 [1,2,24] 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 behavior of formant frequencies with respect to age is showed 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 a study by [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.



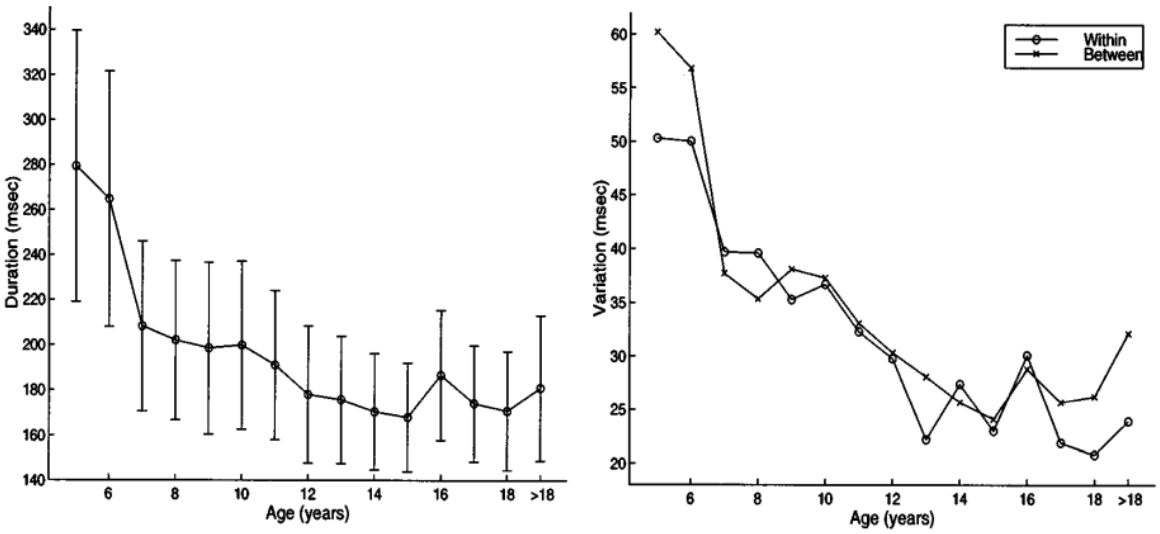
(a) Changes in F1-F2 vowel space as a function of age

(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 exhibits intra-speaker variability, signifying that the speech of the same individual can exhibit variations in different ways. This variability arises from two primary sources. Firstly, as previously discussed, the acoustic characteristics of a same 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 refine 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 [25], is: *"Any discrete unit that can be identified, either physically or auditorily, in the stream of speech"*. These segmental durations could be vowel or sentence duration. Vowel duration, in particular, is of



(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]

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 vowels 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 as they 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 in the context of ASR, where the age of the child is often unknown. In addition, the dynamic processes of vocal tract growth, changes in speech components, and the maturation of speech motor control in children occur simultaneously and overlap, making it considerably more challenging to accurately disentangle and model their acoustic effects. The intricate nature of children's speech, marked by age-related variations in frequency, formants, and segmental duration, 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 uniquely human ability and serves as a fundamental aspect of human cognition and social interaction. Linguists have identified five basic components of language [26],

including phonology (sounds), morphology (struce 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 crucial milestone in their cognitive development, and language plays a central role in shaping culture, identity, and the transmission of knowledge to them. The children's ability to use language effectively develops with age, achieving adult capabilities around the age of 13 , as indicated by research [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 learning process of language acquisition, children, constrained by their limited and developing linguistic knowledge, often make pronunciation errors and encounter disfluencies [27]. According to [28], 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 rendition.
- **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 word or a part of a word incorrectly, deviating from the standard or expected pronunciation in a particular language or dialect.
- **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 an utterance 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 [29] 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 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.

Mispronunciations and disfluencies are not necessarily present in adult speech to the same extent as in children's speech, as supported by studies such as [30,31]. These studies used language models specifically trained on children's speech, demonstrating the advantages over the use of adult language models. The 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 about significant advancements in the ASR field. The combination of increased computational power and the abundance of available data has played a pivotal role in driving these improvements. The success deep learning is largely attributed to the deep neural networks (DNNs), which effectively approximate the complex non-linear functions. With the help of the capability, DNNs excel in capturing intricate patterns and representations in speech data, contributing to more accurate and robust ASR systems. However, the efficacy of a DNN in capturing intricate speech patterns depends a lot on the availability of a substantial amount of training data. The emphasis on accumulating expansive datasets is driven by the recognition that a large of diverse and comprehensive training data is pivotal for enhancing the capabilities and generalization of DNN-based ASR systems. Notably, top-performing ASR systems like Whisper have been trained on extensive datasets, surpassing 680,000 hours of data collected from the web [32]. 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 [33], and the GigaSpeech dataset, featuring a staggering 10,000 hours [34].

Unfortunatly, despite rare recent efforts to collect larger children datasets [35–37], 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 [35, 37–40]. However, English is a large-size, 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 attributed to a combination of ethical, legal, and technical challenges. Collecting speech data from children raises ethical concerns related

to obtaining informed 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 previously, making it 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 a classroom), 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 encompass greater variability has been emphasised in a study conducted by Liao in 2015 [41]. In this work, the researchers addressed the limited availability of extensive children's speech datasets by training an ASR model using a sizable 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, demonstrating competitiveness with adult speech recognition systems. This study underscores the crucial role of comprehensive and diverse children's speech datasets in developing robust and high-performance ASR models tailored to the distinctive 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 [42–44]. The section is structured as follows: Firstly, we present the historical evolution of ASR, followed by an description of traditional 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 in a speaker-dependent context. One of the earliest project in this domain was the creation of a digit recognizer 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, achieving an impressive accuracy of up to 99%. The system relied on formant frequency approximations to recognise entire words. It is impotant 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 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 pattern-matching algorithm similar to the pattern-matching approach used in the Bell Telephone Laboratories's recogniser.

Nevertheless, extending such a system for 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

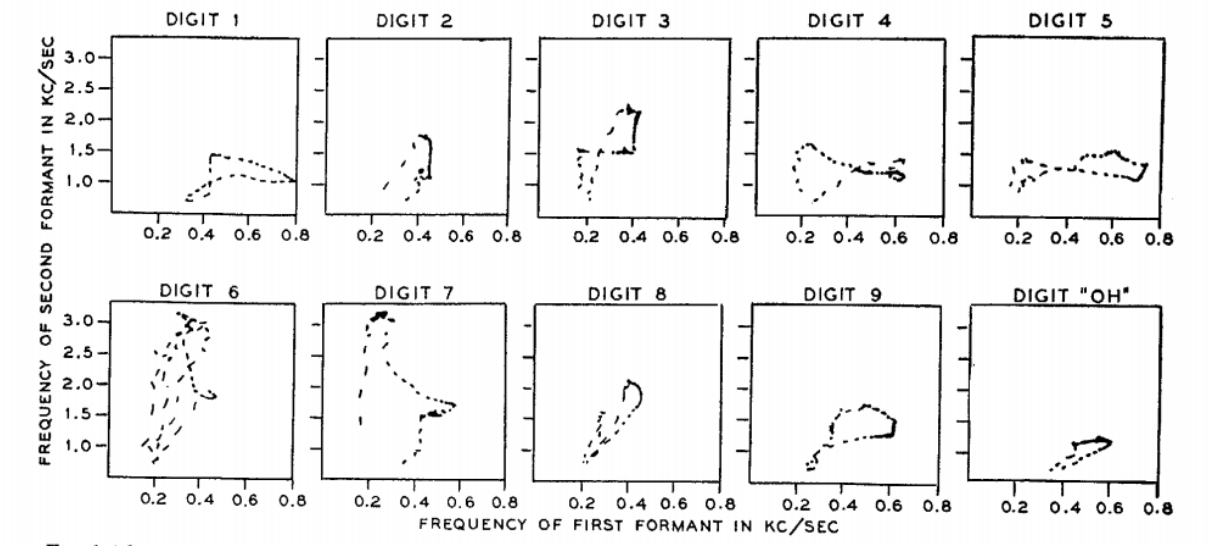


Figure 2.3: Example of a standard digit pattern from Davis et al. 1952

to automatic large vocabulary recognition, issues of time and disk usage complexity emerged as significant challenges. 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 continued 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, from a vocabulary of 1,000 words, 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. In 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 an almost 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. In sentence understanding, the system needed to handle variable sentence lengths, making it necessary to adopt a more flexible approach in modeling and recognizing speech.

To address these challenges, a shift was made to a smaller unit than the word for modeling 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 Carnegie Mellon team exhibited the best performances [3]. Harpy is a speaker-specific system that use 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, digitization 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 exemplar illustration of a decoding graph in the Harpy 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 recognising a vocabulary of no more than 1,000 words and relies on a 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 template-based approaches became apparent. This realisation prompted the exploration of probabilistic modeling techniques, marking a shift towards more sophisticated and adaptable approaches in automatic speech recognition.

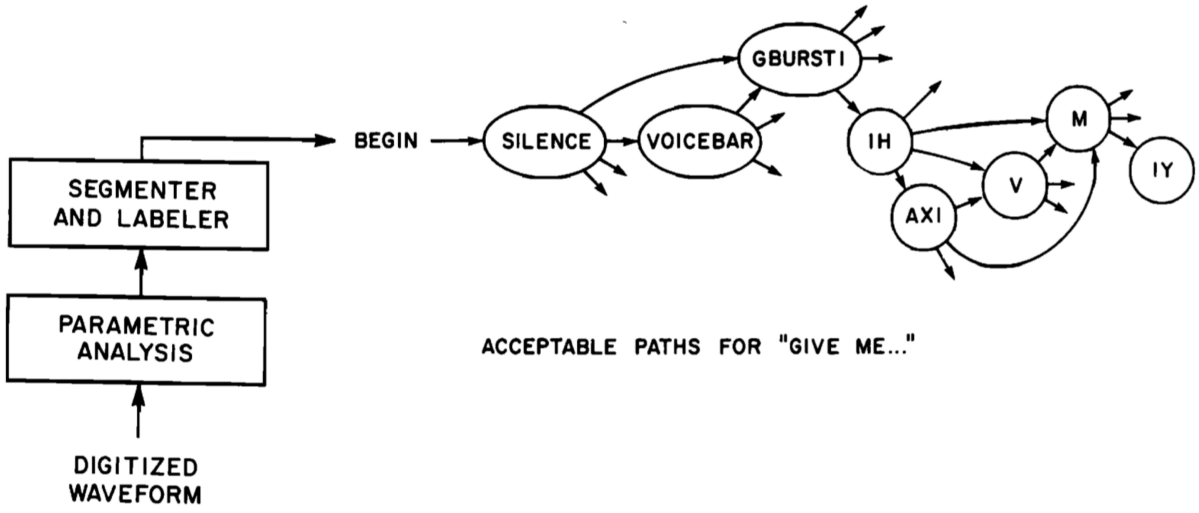


Figure 2.4: Example of a decoding graph from the Harpy system for the sentence "GIVE ME" from [3]

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 [45]. 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. HMMs model the transitions between these states and, at each state, generate 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 modeling capabilities of ASR [46]. 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. They assume that the distribution of features can be approximated by a mixture of several Gaussian distributions. GMMs 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 modeling capabilities of HMMs with the statistical representation power of GMMs, this framework effectively captures the complex relationship between acoustic features and phonetic units.

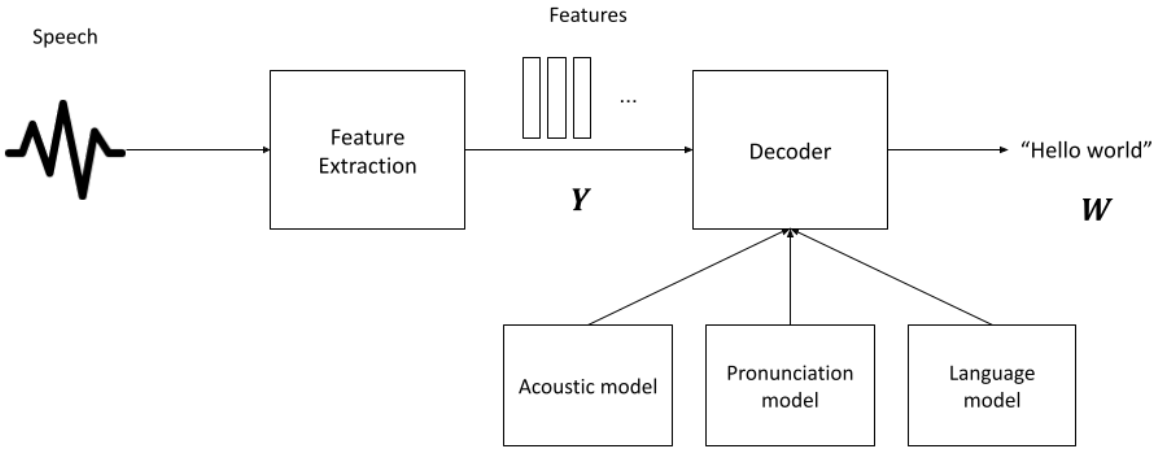


Figure 2.5: Architecture of a HMM-based speech recognition system

Finally, in the 1990s, statistical grammars also played a crucial role, providing a structured framework for incorporating linguistic into ASR systems [47]. 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. These components continue to form the core of modern HMM-based ASR systems. However, in more recent times, the field of ASR has witnessed a transformative shift with the adoption of deep neural networks (DNNs) instead of GMM. Called hybrid models, they combining the strengths of HMMs and DNNs and have further improved the ASR system performance [48]. A DNN is a subtype of artificial neural networks consisting of multiple layers of interconnected neurons. These neurons, organised in layers, receive an input signals, and each connection between neurons is characterised by a weight that signifies its strength. In addition, each neuron is associated with a bias, provide 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, there is a non-linear activation functions 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 Automatic Speech Recognition (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:

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

However, directly modeling 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:

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

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 feature extraction component plays a crucial role in capturing pertinent information about the linguistic content of speech. 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. A acceptable assumption is that speech is considered as 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), mel spectrograms (melspec), filterbanks (fbanks), and Mel-frequency cepstral coefficients (MFCC). However, MFCCs as introduced by Davis and Mermelstein [49], 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 the magnitude spectrum. The next stage involves passing 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, and

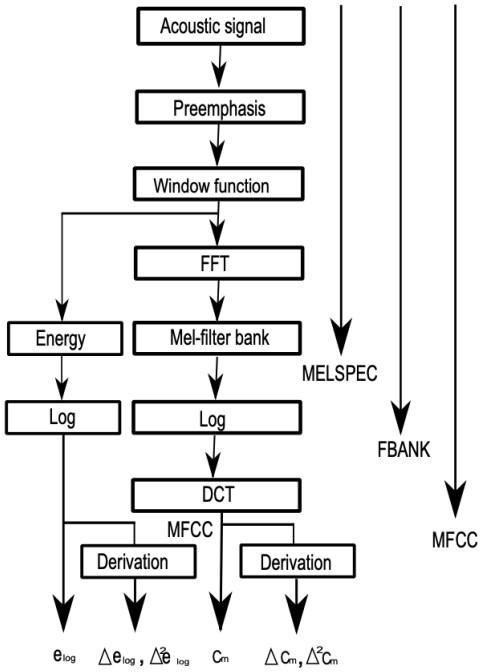


Figure 2.6: Principal block scheme of main speech features for ASR: Melspec, fbanks and MFCC coefficients from [4]

concluding the extraction process at this stage results 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 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 represent 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 neighboring features at $t \pm 2$.The $\Delta\Delta$ coefficients, also written Δ^2, represents the second-order derivatives and are computed in a similar manner as Δin equation 2.3 by taking the difference between 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

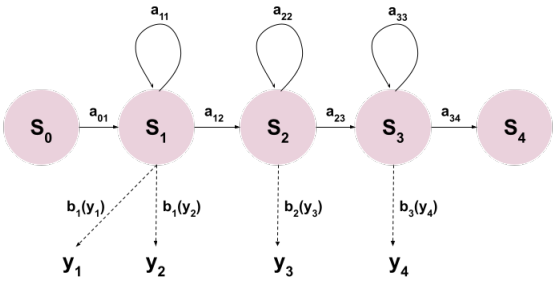


Figure 2.7: Three-state Hidden Markov Model for modelling phones

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, espacially 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 categorizing 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, further complicating the classification task. To address these concerns, the HMM framework has been proposed as a solution [50]. HMMs offer temporal flexibility, incorporating concepts such as self-looping, and provide a well-understood framework with effective learning (Expectation Maximization) 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 modeled by the distribution $b_i(\cdot)$ associated with that state. Typically, this emission distribution is modeled by a GMM. It is assumed that all observations are independent given the states that generated them. A

fundamental configuration in HMMs for speech recognition 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 modeling. For example, since English has 44 phonemes [51], a monophone system on 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 [52]. A triphone system aims to model each phoneme in its specific phonetic context, leading to a significantly larger number of required models. Indeed, or 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 [53]. 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 perceptrons (MLP) as replacements for GMMs in the HMM-GMM system [54, 55]. Subsequently, the introduction of DNNs, which are MLPs with a large number of hidden layers, in 2012 marked a significant advancement in various ASR tasks [48]. 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 labeled data for training, which includes both input features and corresponding output labels (e.g., phoneme labels). Standard speech training data often lacks this detailed labeling, providing only audio waveforms and utterance transcriptions. Consequently, the training of HMM-DNN models relies on alignments generated by an HMM-GMM. Therefore, the training process involves initial flat-start monophone training with HMM-GMM, followed by iterative steps into triphone training with more precise alignments. 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) [56], Long Short-Term Memory networks (LSTMs) [57], and Time-Delay Neural Networks (TDNNs) [58], further refines the modeling of spatial and temporal relationships, laying the foundation for more sophisticated and context-aware speech recognition systems.

2.2.2.C Pronunciation model

The pronunciation model in Automatic Speech Recognition (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. In ASR systems, words are essentially comprised of phonetic segments, and the pronunciation model specifies how these segments combine to articulate the pronun-

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]

ciation of each word. 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 Figure 2.8. Traditionally, this mapping is obtained manually, relying on phonetic and linguistic knowledge. It's noteworthy that a single word may have multiple pronunciations.

Furthermore, the integration of statistical grapheme-to-phoneme (G2P) tools [59] 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 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 extends into diverse fields including Natural Language Processing (NLP) [60], computational biology [61], and data compression [62]. The two most successful approaches to language modeling widely adopted in ASR are, respectively, statistical methods and models based on deep learning.

Statistical language models rely on traditional techniques like HMM and n-grams. N-grams, which are the simplest approach for language modelling, 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 granularity of context varies from the case when $n = 1$, the 1-gram -or unigram (considering 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 modeling, 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. [63].

In contrast, deep learning-based language models have opened up a new era, employing neural networks with complex architectures to achieve remarkable modeling capabilities. Unlike traditional n-grams, these models demonstrate a high degree of flexibility, ease training and do not require as many resources as n-grams to be efficient. Recent advances in language modeling, exemplified by state-of-the-art models such as Bidirectional Encoder Representations from Transformers (BERT) [64] or Generative Pre-trained Transformer (GPT) [65], 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 refered in equation 2.1). This is achieved by employing dynamic programming to search through all potential sequences. Notably, the Viterbi algorithm [66], are instrumental in efficiently solving this decoding problem. However, in practical applications, a direct implementation of the Viterbi algorithm becomes 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 [67] or dynamically expanding it as the search progresses [68]. Another alternative is to use beam search where the idea is to prune search path which are unlikely to successed. 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 [69, 70]. This approach provides both flexibility and efficiency, making it valuable for both ASR research and practical applications. As demonstrated by the Kaldi speech recognition toolkit [71], which 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 equatio 2.1, they can also generate the N-best set of 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 [72] 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.

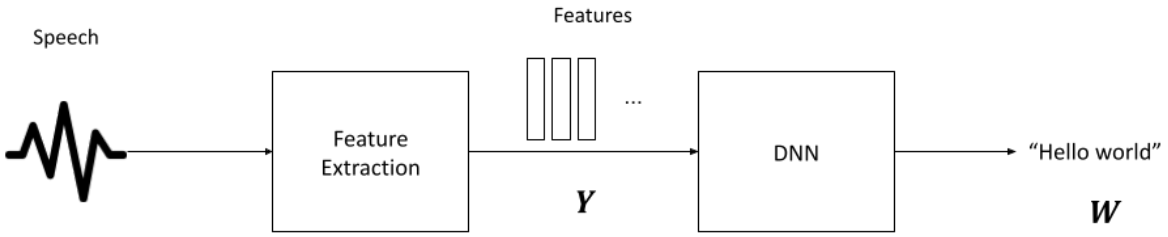


Figure 2.9: Architecture of an end-to-end speech recognition system

2.2.3 End-to-end automatic speech recognition

End-to-end speech recognition signifies a transformative paradigm 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 propose to simplify the ASR process by directly map 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 factorized 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 increased interest in the end-to-end paradigm, these systems has demonstrated notable success, particularly in scenarios where substantial labeled training data is available, leading to competitive performance with traditional HMM-based systems across diverse ASR datasets [73, 74]. However, despite its promising capability, end-to-end speech recognition faces challenges that merit ongoing research and development efforts. Indeed, it is essential to underscore that while these end-to-end systems showcase remarkable capabilities, they operate on a fundamentally different paradigm—being data-driven—and thus necessitate a substantial corpus of training data to fully work correctly [75]. In addition, handling rare or out-of-vocabulary words remains a concern, and the generalisation of models across varying acoustic conditions can be challenging. As research and development in end-to-end speech recognition continue, this approach holds significant potential to simplify the design, training, and

deployment of speech recognition systems, ushering in a new era of efficiency and adaptability.

Two distinct paradigms have stand-out as pivotal in the end-to-end ASR literature: Connectionist Temporal Classification (CTC) and sequence-to-sequence-based architectures. Each approach carries its unique characteristics and advantages, and the subsequent sections provide a more comprehensive exploration of these two methodologies. While these are two distinct methods that can be employed independently, it is also feasible to use them in conjunction.

2.2.3.A Connectionist Temporal Classification

The first step towards end-to-end automatic speech recognition was made with the introduction of the Connectionist Temporal Classification (CTC) objective function by Grave et al. [76]. This novel approach departed from traditional HMM-based models and brought about distinctive advantages. The core innovation lies in the elimination of the need for pre-segmented training data, allowing the model to automatically learn alignments between the N input speech frame $\boldsymbol{X}$ and the output sequence of T phones $\boldsymbol{Y}$ if $N \leq T$.

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 marginalizing 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" becomes "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 [77] and stands as a important paradigm of end-to-end ASR systems. The original context of its application was for machine translation tasks where a 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 [78], has showcased remarkable versatility, extending its efficacy across diverse applications such as image captioning [79], conversational modeling [80], text summarisation [81], and ASR [82].

The core components of a standard Seq2seq model consist of an encoder and a decoder. 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 maximizing the conditional probability of generating the target sequence given the input sequence using negative log-likelihood loss (NLL) or cross-entropy loss (CE).

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 tasks, like 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, Word Error Rate (WER) is the most commonly used for assessing speech-to-text 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 [83] and the French subset of the read speech Common voice dataset [84] respectively. However,it is important to note that WER scores can be task-specific, and 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 [85].

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) for languages based on phonemes and Character Error Rate (CER) which operates on character instead of word. 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 speech patterns. Modifications in annotation detail have been explored, refining the labeling 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 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 phases of the thesis.

2.3.1 Feature extraction and adaptation

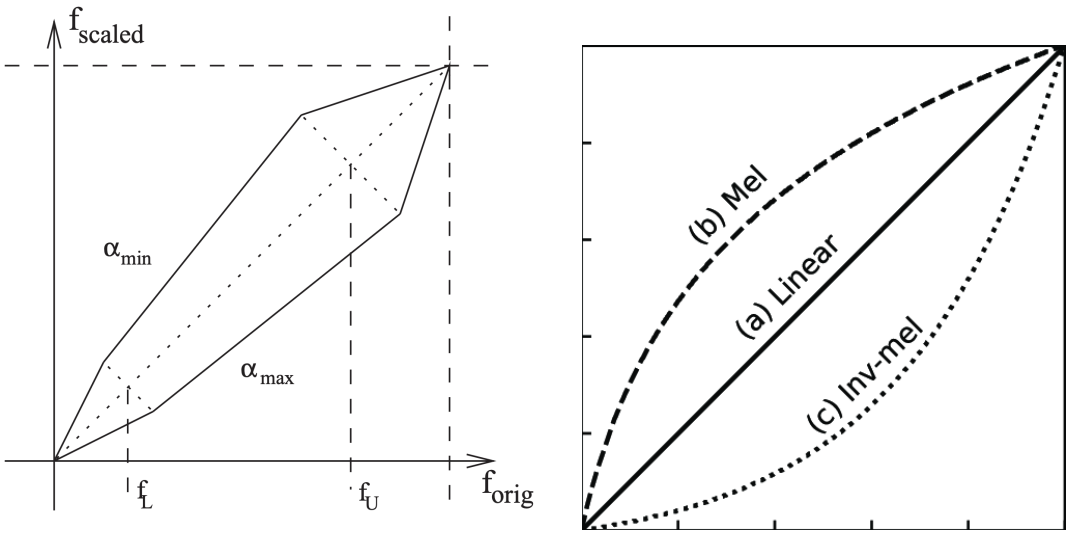
The feature extraction stage is critical for identifying relevant speech signal components for automatic speech recognition. Achieved by discarding speaker-dependent information such as fundamental frequency while retaining phoneme-dependent information such as formant frequencies. However, the acoustic fluctuation of children's speech leads to close fundamental frequency and formant values, as well as phonetic class overlap due to formant value variability, making typical short-term spectral-based feature extraction challenging. Several strategies have therefore been proposed to improve children's acoustic features, either directly at the extraction stage or at the feature level.

2.3.1.A Feature extraction

Feature extraction methods aim to extract acoustic features directly from the raw speech signal that are capable of better suit children's speech characteristics. The initial step toward this direction was the introduction of perceptual linear predictive (PLP) features in 1990 by [86], where PLP features demonstrated a more accurate formant representation of children's speech. More recently, [87] proposed to shift the feature extraction stage from a hand-crafted fashion to data-driven with a feature learning strategy directly from the raw signal. The intuition behind this end-to-end approach is that hand-crafted features rely on adult speech analysis, whereas children's speech has greater acoustic variability and may not be well suited. In this study, data-driven feature extraction using Convolution-neural-network (CNN)

based models discovered relevant features that yield better results than the standard hand-crafted ones. A similar idea has been proposed by [88] in which the convolutional layer of the feature extractor has been replaced by a SincNet layer [89] to effectively adapt an adult end-to-end features extractor to children.

2.3.1.B Feature adaptation



(a) Piecewise linear VTLN frequency scaling function for warp factors α. f_L and f_U denote lower and upper threshold frequencies [90]

(b) VTLN frequency warping functions for a) linear, b) mel, and c) inverse mel frequency scales [91]

Figure 2.10: Examples of VTLN frequency warping functions

Another perspective to reduce children's acoustic variability is to work directly at the feature level, using standard feature extractors and directly adapting speech features to reduce children's acoustic variability. To this end, Vocal Track Length Normalization (VTLN) has been widely used to wrap spectral features into a canonical space [90, 92]. Applied during the feature extraction process, after the windowing and Fourier transform, VTLN stretch or compress the frequency axis of the spectrum depending on a warping function (several examples of warping function are presented in figure 2.10). As a result, the spectral representation of speaker variance is reduced, and children's speech features are normalised. In addition, adapting acoustic models with Maximum Likelihood Linear Regression (MLLR) or Maximum A-Posteriori (MAP) and using Speaker adaptive training (SAT) based on feature MLLR (fMLLR) or Constrained MLLR (CMLLR) were found to be effective to improve the performance of children ASR [1, 31, 93, 94].

Recently, some research explored normalising several aspects of speech: Pitch to reduce the spectral mismatch between children [95–97], formant values to better match adult ones [98] or speaking rate with a time-scale modification approach [99].

Moreover, in accordance with the trend of shifting from knowledge to data drive framework, some recent studies attempted to use data-based feature adaptation. For example, [100, 101] used adversarial multi-task learning to produce age-invariant features which minimize the acoustic mismatch between adult and children's speech.

2.3.1.C Additional features

In addition to feature extraction and feature-level adaptation, some research has demonstrated that concatenating speaker embeddings, such as i-vectors, to acoustic characteristics leads to more speaker-independent models [102]. Similarly, [103] proposes concatenating various prosodic variables such as loudness, voice intensity, and voice-probability to standard acoustic features, resulting in decreased inter-speaker variances and enhanced phoneme class discrimination.

2.3.2 Detail of the annotation

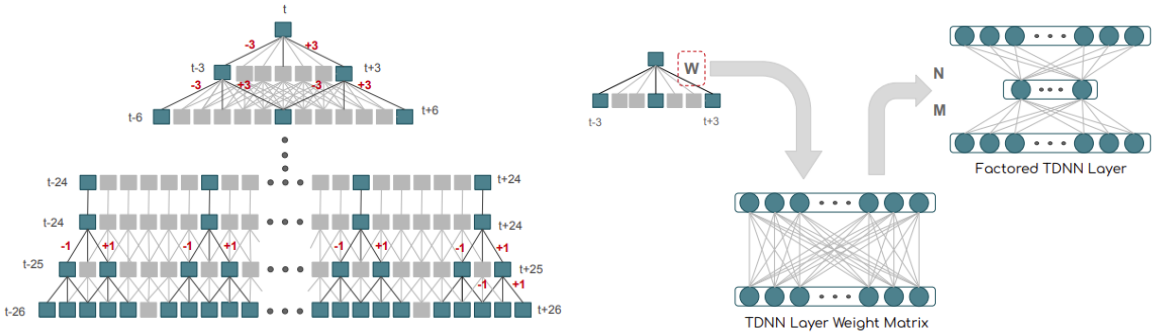
The material provided in speech corpora is frequently restricted to the audio signal, the corresponding text transcription, and some anonymised speaker's identifier. However, some additional information might be useful in improving children's speech recognition. The incorporation of the speaker's age would allow the development of age-dependent ASR systems in which the acoustical fluctuation caused by vocal track growth would be reduced [104]. Some work showed the efficiency of an annotation at the sub-word level instead of the word level [105] especially to be more robust to mistakes such as mispronunciations or hesitations. Alternatively, in [93, 106] user-dependent pronunciation lexicons are employed to tackle the pronunciations divergence from the canonical and adult patterns.

2.3.3 Structure of the acoustic model

The acoustic model is a key element in the recognition of children's speech since acoustic variability plays a major role in the degradation of children's speech recognition (compared to linguistic variability). As a result, the structure of the acoustic model is essential to be more robust to children's speech variability. In the rest of this section, we will review children's acoustic modelling structures in the context of hybrid models and end-to-end systems.

2.3.3.A Hybrid models

Acoustic models for children generally follow the latest advances in acoustic modelling for adults. For example, [8] proposed switching from GMM to DNN for hybrid models. However, the traditional fully connected neural network does not provide the contextual information required for efficient phoneme recognition. Thus, time-delayed neural networks (TDNN) have been proposed ([107]). Working similarly to a one-dimensional convolutional neural network, where at each time step the current time step and its



(a) **(b)**

Figure 2.11: (a) TDNN layers with sub-sampling & (b) Factorized TDNN layer from [6]

corresponding left and right context are provided, as shown in figure (2.11(a)). Additionally, the use of multiple layers of TDNN allows for the capture of broader contextual information, which leads to better phoneme recognition. More recently, factorized time delay neural networks (TDNN-F) were introduced as a TDNN improvement [6] and proved to be effective for children ASR [108]. They differ from TDNNs by applying a singular value decomposition (SVD) on the weight matrices W, which factorize them into two smaller rank matrices:

$$W = U\Sigma V^T = MN \tag{2.12}$$

Where U is a $m \times m$ complex unitary matrix, Σ a $m \times n$ non-negative rectangular diagonal matrix, V a $n \times n$ complex unitary matrix, M a $m \times k$ real matrix and N a $k \times n$ real matrix. With a suitable value of k (much small than m and n), this decomposition preserves the capabilities of the model, reduces the number of parameters and acts as a bottleneck layer (as shown in figure 2.11(b))

2.3.3.B End-to-end models

More recently, the ability of end-to-end models to outperform the hybrid HMM-DNN models in terms of absolute WER for children has been demonstrated [109–112]. Nonetheless, a fundamental issue for this type of model is the requirement for a large amount of data to function correctly. Therefore, in most cases, training reliable end-to-end model for children from scratch is difficult. As a result, different training approaches are required, such as transfer learning (which will be detailed in further detail in section 2.3.5.A) or semi and self-supervised learning, when there are few or no transcriptions, respectively [113].

2.3.4 Data augmentation

Deep learning's success is mostly due to its capacity to effectively utilise massive amounts of data to recognise patterns and be robust to variances. As a result, the lack of data on children's speech plays a significant part in performance deterioration as compared to adults. Data scarcity for children's speech

is especially problematic for languages other than English, for which fewer resources are accessible in general. As a consequence, the need for large amounts of data has inspired research into a variety of data augmentation approaches, the goal of which is to artificially increase the amount of data for training. There is two way of conducting this augmentation, according to the literature: using solely the data that is currently accessible or incorporating external data.

2.3.4.A Using external data

The use of out-of-domain adult data has been shown to be effective in improving the automatic recognition of children's speech [114, 115]. In particular, improvements were observed using adult female speech, as the acoustic mismatch between females and children is smaller than adult men. Similarly, [116] proposed to augment using directly additional children data.

Furthermore, several studies proposed leveraging new data generated by deep learning systems as additional data during training. This is done in the literature using two families of algorithms: generative adversarial network (GAN) and text-to-speech (TTS).

Generative Adversarial Networks are generative models that generate new data instances that are similar to the data on which they were trained. For that purpose, the GAN model is separated into two parts: a generator, which generates the synthetic data, and a discriminator, which receives either the created or actual data sample and attempts to differentiate between the two. During training, each module competes with the other, with the generator attempting to trick the discriminator and the discriminator attempting not to be fooled. A GAN was employed in [117] to transform children's speech to adult speech in order to reduce variability. This was accomplished by feeding the generator children's utterances and utilising adult data as real data. In this approach, the discriminator seeks to differentiate between altered children's speech and real adult speech. At inference time, the discriminator is removed, leaving just the generator to turn children's speech into adult speech.

Text-to-speech systems aim to generate speech examples using directly the transcription as input. With recent advance of TTS systems such as Tacotron2 [118] or VITS [119], more realistic speech utterances are produced. Naturally, some research tried to use TTS outputs as data augmentation for ASR task [120] since the transcription and speech data are available. However, children's speech contains more complex traits than adults such as substandard or unclear pronunciation in addition to acoustics variability. As a result children's TTS quality is often inconsistent. In consequence, [121] proposed different data selection strategies such as speaker embedding similarity between the reference speaker and the speaker embedding extracted from generated speech utterance. With this data selection strategy, they have significantly reduced the CER for different children's speech recognition tasks.

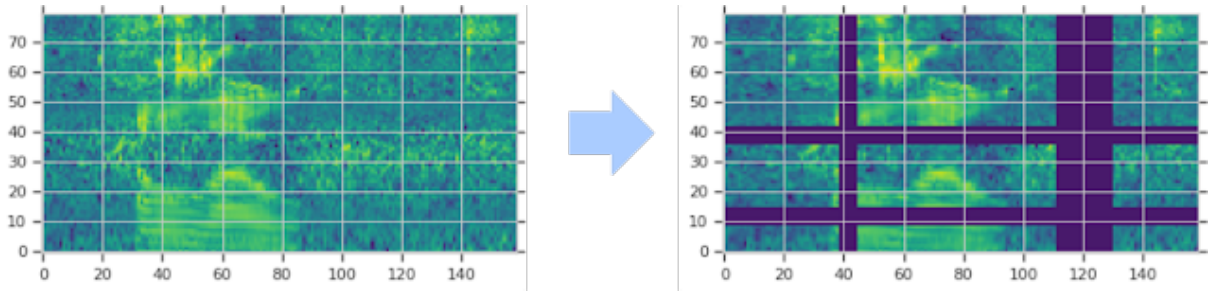


Figure 2.12: Before-After specaugment augmentation with warping of the time and time steps and Mel frequency masking (figure from [7])

2.3.4.B Using available data

A common procedure for enhancing the robustness of the model by leveraging existing data is to construct a copy of the original data with different noises (such as babel noise, white noise, and music) and reverberation added [122–126]. More recently, vocal tract length perturbation (VTLP) [127] and spectral augmentation (SpecAugment) [7] became popular data augmentation techniques, in particular for end-to-end systems. They create a new copy of data by warping the frequency axis. SpecAugment also randomly masks time and frequency bands of the original audio at the feature level as shown in figure 2.12.

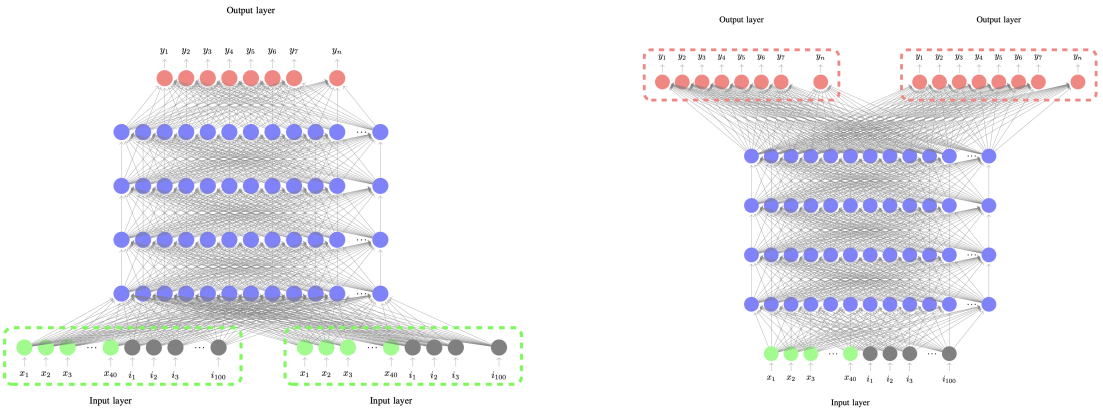
Finally, as mentioned in section 2.1.2, children's speech contains many disfluencies and errors which can complicate the learning process of the ASR model. Therefore, to make the model more robust to such errors, [128] proposed to synthetically create manual reading errors, such as word insertion or substitution by cutting the signal or adding speech elements produced by other children.

2.3.5 Training procedure for children speech recognition

The most typical method for training a deep learning system is to alternate between forward and backward passes. The corpus input is supplied to the network in the forward pass to create a prediction. The loss is calculated using both predicted and actual target values. To reduce the number of prediction errors, network weights are adjusted using the gradient descent technique in the backpropagation phase. The model is considered trained after a number of loops between these two passes. There are, however, variations of this kind of pipeline in the literature such as transfer and multi-task learning, which stand out in the context of children's voice recognition.

2.3.5.A Transfer learning

When presented with a new problem in biological intelligence, people can employ information from a previous task as an inductive bias. Humans may avoid having to learn everything from scratch by transferring knowledge. It may be defined as the capacity to identify and utilise past task knowledge as



(a) Acoustic adaptation **(b)** Pronunciation adaptation

Figure 2.13: Transfer learning approaches. Figures from [8]

a starting point for new tasks. In contrast, algorithms in machine learning are generally developed from scratch on a specific task. Transfer learning (TL) or parameters transfer is a concept that has emerged to bridge the gap between artificial and biological intelligence. To do this, the model's parameters are initialised with parameters obtained from another well-resourced model trained on a source-related task. Then, this model is adapted with the data from the new domain by adjusting the parameters to better fit the target task. The resulting model leverages various underlying characteristics that have been captured by the different layers of the neural network during the training on the source domain step. A common assumption in deep learning is that the bottom layers, closest to the input, capture more signal-specific characteristics. While higher layers, near the output, capture task-specific information [129, 130]. Furthermore, because the target model is based on a pre-trained model, one advantage of TL is that it requires less training data for adaptation.

In recent years, TL has been successfully used in a wide variety of applications, especially for low-resource tasks, such as language understanding [64], character recognition [131] and dysarthric speech recognition [132] among others. These successes have motivated its use for children's speech recognition. In particular, using large adults corpora as source domain. Indeed, recent acoustic models trained on adults are increasingly efficient and contain many acoustic and phonetic information that can be used for efficient adaptation to children's speech. Because children's speech variabilities are in both acoustics and pronunciation, [8] proposed to study three different methods of TL to respectively access the contribution of acoustic adaptation, pronunciation adaptation and the combination of the two.

Acoustic adaptation is based on the well-established idea that lower-level layers capture acoustic properties. In consequence, to solve children's acoustic variability that affects the lower-level layers, acoustic adaptation consists of freezing the top-level layer's weights and using TL on the lower-level

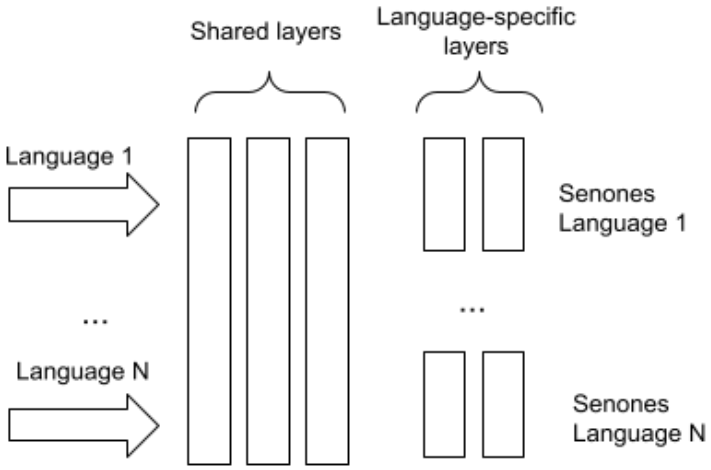


Figure 2.14: Multilingual approach using each language as a task in a multi-task learning context.

layers as shown in figure 2.13(a). For [8] and [133], acoustic transfer learning from an adult model for children, by retraining only the first layers of the model led to respectively 38% and 26% WER relative improvements compared to adult models performances. Furthermore, with a 4.9% relative WER improvement, the children's acoustic adaption outperforms a randomly initialised acoustic model trained on the same children's data.

Pronunciation adaptation is based on the assumption that higher-level layers capture task characteristic information, which in the context of ASR is pronunciation. Pronunciation is the act or manner of pronouncing words using phonetic symbols, phonetic symbols which correspond to the different outputs of the acoustic model. Pronunciation adaptation, as illustrated in figure 2.13(b), adapt higher layer, keeping lower-level layers weight frozen. In [8], pronunciation adaptation on the last layers of the model led to 31% relative WER improvement compared to the adult model's performance. However, as compared to a randomly initialised acoustic model trained with the same children's data, pronunciation adaptation degrades by 5.6% relative WER.

Finally, fine-tuning the entire network outperforms both acoustic and pronunciation adaptation alone. Similar findings have recently been seen in end-to-end models, where TL leveraging an adult pre-trained model outperformed training from scratch using only children's data [109, 110].

2.3.5.B Multi-task learning

Multi-task learning (MTL), like transfer learning, is inspired by biological intelligence. By training all tasks in simultaneously, MTL tries to discover shared representations between related tasks. MTL differs from TL in that it does not limit learning to only the source and target tasks in sequential training, but

instead permits learning with as many tasks as possible in parallel. In general, a typical MTL model consists of two different parts. The first part is the sub-network shared by all tasks, while the second is a task-specific output sub-network (see figure 2.14). The joint representation learned in the shared layers is more robust, allowing the model to be more reliable. More formally, for any task i the corresponding output of the forward pass will be:

$$f(X_i; \{M_i, M_c\}) = f_i(f_c(X_i, M_c); M_i) \tag{2.13}$$

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.

Therefore, the performances of the MTL greatly depends on the task-relatedness. Indeed, the MTL is sensitive to outlier tasks that are unrelated to the rest of the tasks. This is due to the difficulty to learn common representations for tasks that are unrelated to each other [134].

MTL has been used effectively in a variety of areas, including natural language processing [135], computer vision [136] and bioinformatics [137]. MTL has been naturally applied in the field of automatic speech recognition [138] with directly application to low-resource ASR [139]. Given that child ASR is a resource-limited task, MTL has been proposed to address the issue of data scarcity by utilising various datasets of children's speech. [133] and [140] successfully applied MTL to Mandarin and English-speaking children, with a 16.96% relative improvement in WER for the English children.

2.4 Children Corpora

As described in chapter 2.1.3, notwithstanding recent efforts to collect databases on children's speech, there are still fewer data available than for adults. In addition, collecting speech data for children is challenging in many ways, from the child's attention span, mispronunciations, and ungrammatical expression to the use of out-of-ordinary words. Therefore, the scarcity of publicly available child speech corpora and their small size impeded research and further development of reliable ASR systems for children.

Table 2.1 provides a list of existing corpora of children's speech. Notably, one-third of the available corpora are in English. Similarly, one-third is focused to children under the age of 4. Furthermore, there is always a trade-off between speaker diversity, total duration, and utterance number.

The remainder of this section will go through the children's speech corpora that were used in this proposal.

Corpus	Languages	# Spkrs	# Utt	Dur.	Age Range	Date
Providence Corpus [141]	English	6		363h	1-3	2006
Lyon Corpus [142]	French	4		185h	1-3	2007
CASS_CHILD [143]	Mandarin	23		631h	1-4	2012
Demuth Sesotho Corpus [144]	Sesotho	4	13250	98h	2-4	1992
NITK Kids' Speech Corpus [145]	Kannada	160		10h	2-6	2019
CHIEDE [146]	Spanish	59	15,444	8h	3-6	2008
CUChild [147]	Cantonese	1,986			3-6	2020
EmoChildRu [148]	Russian	100	20,000	30h	3-7	2015
CNG Portuguese children [149]	Portuguese	510		21h	3-10	2013
AusKidTalk[1] [37]	English	750		600h	3-12	2021
PF-STAR-SWEDISH [150]	Swedish	198	8,909	6h	4-8	2005
PF-STAR Children British [40, 150, 151]	English	158		14.5h	4-14	2006
AD-child. RU [152]	Russian	278			4-16	2019
TBALL [153]	English	256	5,000	40h	5-8	2005
SPECO [154]	Hungarian	72		12h	5-11	1999
UltraSuite [155]	English	86	14,456	37h	5-14	2019
CID read speech corpus [156]	English	436			5-18	1996
Persian Kids Speech Corpus [157]	Persian	286	162,395	33h	6-9	2022
Letsread[2] [158]	Portuguese	284	4,629	14h	6-10	2016
CMU kids Corpus [38]	English	76	5,180		6-11	1997
CFSC [159]	Filipino	57		8h	6-11	2012
IESC-Child [160]	Spanish	174	19,793	34h	6-11	2020
CU Children's read and prompted [161]	English	663	66300		K-G5	2001
Chorec[2] [162]	Dutch	400	3,065	25h	6-12	2008
ChildIt2 [163]	Italian	96	4,875	9h	6-14	2016
TIDIGITS [164]	English	101			6-15	1993
CSLU Kids' Speech Corpus [39]	English	1,100	1,017		K-G10	2007
SingaKids-Mandarin [36]	Mandarin	255	79,843	125h	7-12	2016
ChildIt corpus [165]	Italian	171			7-13	2007
VoiceClass Database [166]	German	170			7-14	2010
Deutsche Telekom telephone [166]	German	106			7-14	2010
Jasmin [167]	Dutch			63h	7-16	2008
Tgr-child corpus [165]	Italian	30			8-12	2007
SponIt corpus [165]	Italian	21			8-12	2007
Swedish NICE Corpus [168]	Swedish	75	5,580		8-15	2005
CHIMP spontaneous speech [29]	English	160			8-14	2002
SpeeCon corpus [169]	20 Languages				8-15	2002
Rafael.0 telephone corpus [104]	Danish	306			8-18	1996
Boulder Learning - MyST [35]	English	1,371	228,874	384h	G3-G5	2019
CU Story Corpus [161]	English	106	5,000	40h	G3-G5	2003
ETLT[2] [170]	L2 German		1,674	6h	9-16	2020
Lesetest corpus [171]	German	62			10-12	2000
FAU Aibo Emotion Corpus [172]	German	51	13,642	9h	10-13	2002
PIXIE corpus [173]	Swedish	2,885				2003
Takemaru-kun corpus [174]	Japanese	17,392				2007
CALL-SLT [175]	German		5,000			2014

[1] At the day of this proposal, 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 PhD thesis proposal. K: Kindergarden. G: Grade

2.4.1 LETSREAD

LetsRead database [158] 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.

2.4.2 PFSTAR_SWEDISH

The PFStar children's speech corpus [150] 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 [170] 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 proposal, we only used the German-transcribed subset (named ETLTDE), around 6h divided into training and test partitions.

2.4.4 CMU_KIDS

The CMU kids corpus [38] 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 [176] automatic speech recognition system within the LISTEN project at Carnegie Mellon University (CMU).

2.4.5 CHOREC

The Chorec corpus [162] 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 proposal, similarly to the LETSREAD dataset, we discarded pseudo-word utterances.

2.4.6 MyST

My Science Tutor (MyST) Children Speech Corpus [35] 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 third, fourth and fifth grades. Partitioning of the corpus is already available, ensuring reasonably 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 this proposal, we decided to remove all utterances shorter than one second and longer than 30 seconds. After this filtering, 81971 utterances from 736 speakers for a total of 151 hours remain.

2.5 Summary

In this chapter, we provided a brief overview of children's automatic speech recognition, its challenges, and the research community's response. Indeed, ASR in children has long been acknowledged as challenging. This is mostly due to the development of the vocal apparatus, which produces an acoustic gap with adult speech; nonetheless, this mismatch diminishes until the age of 15. Despite all efforts, it remains an active and challenging area of research. In this chapter, we also discussed inductive bias approaches, such as transfer learning and multi-task learning, which would be applied in the work detailed in this proposal. Furthermore, we present the most complete comparison of children's speech corpora to our knowledge.

3

Hybrid models for children automatic speech recognition

Contents

3.1 Introduction

In Chapter 2, we saw how a HMM-GMM or HMM-DNN ASR system may integrate knowledge into the automatic speech recognition pipeline. It uses knowledge from language model, acoustic model, and vocabulary to reduce the amount of speech data required to generate appropriate results. According to the literature, these results can be further improved with inductive bias approaches such as transfer learning and multi-task learning [133]. For years, hybrid configurations have been a privileged setting for the children's ASR community. As a matter of fact, between 2009 and 2020, 80% of published research on children's speech recognition was based on hybrid systems, with 45% using HMM-GMM and 35% HMM-DNN. However, during the same period, 63% of published work was conducted for English [177]. As a result, it is uncertain how children's speech from other languages relates to the various approaches used in English, particularly transfer and multi-task learning.

This chapter will investigate these strategies in a variety of scenarios employing non-English data. First, we present transfer and multi-task learning using adult speech as an inductive bias, as is common in the literature. Second, because there is a lack of data for both adult and children's low-resource languages, we investigate the same methodologies using only children's speech. Finally, we present our approach, multilingual transfer learning, which combines transfer and multi-task learning to produce a more robust model for speech recognition in low-resource children setting.

3.2 Multi-task and Transfer learning using adult and children data

3.2.1 Methodology

Motivated by the success of knowledge transfer approaches for ASR children using adult data in the research [8, 133, 140], we intend to validate these findings using a low-resource language. Indeed, using adult data for pre-training makes sense since adult speech is more stable and less prone to variation. Using adult speech to train a speech recognition algorithm makes it simpler to extract and recognize intrinsic and meaningful speech patterns.

For this proposal, we assess children's speech recognition performances in four distinct configurations:

1. **Adult model**: Using a model trained from scratch with only adult data.

2. **Children model**: Using a model trained from scratch with only children data.

3. **Multi-task model**: Using a model trained jointly on adult and children data in parallel using multi-task learning.

4. **Transfer learning**: Using a model that has been fine-tuned on children data from the adult model of configuration 1.

3.2.2 Corpus

As stated in the introduction, we aim to evaluate the performance of children's speech in a low-resource language. To this end, we decided to use European Portuguese corpora. European Portuguese can be considered a low-resource language since most adult speech corpora do not exceed 100 hours [178]. In this experiment, we used LetsRead, a child corpus, described in section 2.4 and BD-PUBLICO as adult corpus. The statistics of all these two corpora are provided in the following table 3.1. The rest of this section provides further information about the BD-PUBLICO corpus.

Corpus name	Train	Test
BD-PUBLICO	8085 utt	412 utt
Adult	21h48	01h10
LETSREAD	3590 utt	1039 utt
Children	12h00	02h30

Table 3.1: Number of utterances and duration of the different corpora for multi-task and transfer learning experiments using adult and children data

BD-PUBLICO

The BD-PUBLICO database (Base de Dados em Português eUropeu, vocaBulário Largo, Independente do orador e fala COntínua) [179] consists of reading sentences extracted from Portuguese newspaper PÚBLICO. The sentences that are read correspond to a total of 6 months of news (equivalent to 10M words and 156k different forms). It is composed of 120 speakers, and graduate and undergraduate students from Instituto Superior Técnico (Lisbon). This corpus is considered an adult dataset since all students are between 19 and 28 years old. All recordings were performed in good noise condition, in a soundproof room at INESC-ID (Lisbon), at a sampling frequency of 16kHz and using a high-quality microphone. In addition, a pronunciation lexicon with citation phonemic transcriptions for each word was produced. Finally, manually corrections were applied to the automatically generated transcriptions.

We divided the BD-PUBLICO corpus into three unique sets with balanced gender partitioning: 1) A training set of 80 sentences by 100 speakers. 2) A development set of 40 sentences performed by a total of 10 speakers. Finally, a test set of 40 sentences by 10 speakers.

3.2.3 Experimental setup

All experiments were carried out using the Kaldi open-source toolkit [71]. First, for each corpus, an independent HMM-GMM acoustic model was trained to produce the necessary alignment for the HMM-

DNN model. Then, HMM-DNN acoustic models were trained using 40-dim filter-banks (fbanks) in addition to a 40-dim Spectral Subband Centroid (SSC) features [180]. These features are known to have similar properties to formant frequencies. Thus, we expect them to help vowel recognition and lead to better recognition of children's speech. The resulting 80-dim input features are then augmented by a 100-dim i-vector. Concatenating speaker embeddings to the input features helps to improve model speaker robustness [102]. For our experiments, we use an i-vector extractor trained on a set of pooled children data from different languages.

Data augmentation was applied to all training corpora by perturbing the speaking rate of each training utterance by 0.9 and 1.1 factors; as well as volume perturbation. This helps the network to be more robust to rate and volume variability on the test sets. To further improve the robustness of the model, Specaugment [7] was applied on top of the fbanks and SSC features by randomly masking time and frequency bands.

For all experiments, we kept the same HMM-DNN acoustic model architecture using lattice-free maximum mutual information (LF-MMI) objective with a learning rate of 2.0E-4. The acoustic model architecture is divided into two parts: i) six convolutional neural network layers and seven TDNN-F layers of dimension 1024 and followed by ii) two TDNN layers of dimension 450 and a fully-connected layer.

For the transfer learning experiments, only the first part of the network will be fine-tuned, while the second part will be dropped and replaced by randomly initialized ones. Similarly, for the multi-task learning experiment, the first part will be shared between the adult and the child, while the second part will be independent.

3.2.4 Results

Method	Adult WER ↓	Children WER ↓
Adult model	3.82%	102.83%
Children model	45.56%	26.88%
Multi-task model	4.59%	27.65%
Transfer learning	-	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 notice that employing a model trained on adult data yields a WER of 102.83% on the children's test set. This model achieves 3.82% for BD-PUBLICO. This degradation in the children's compared to adults' scores demonstrates the presence of considerable variability in children's speech, which has a detrimental impact on the ASR scores. It supports the idea that an acoustic model designed exclusively for children is necessary because child speech is currently unusable with adult systems.

Training the acoustic model directly on the children's data, on the other hand, considerably improved the word error rate on the children's data to 26.88%. Since the model observes acoustic variability during training, it becomes more robust to it. While the model improved for children, it deteriorated adult speech recognition performance to 45.56 % WER. This confirms the acoustic mismatch between adult and children speech once more. We compare transfer and multi-task learning approaches using these two experiments as a baseline.

When the model was trained jointly utilising adult and children data in the scenario of multi-task learning, the recognition score of the adults and children decreased marginally when compared to the adult and child model baselines. Unlike in the adult and child models, where the mismatch significantly reduced the children's score in the adult model and the adult's score in the child model, both recognition scores in this multi-task learning scenario are comparable to their respective "trained from scratch" baselines. These results were achieved by including corpus-specific layers into the acoustic model architecture. Indeed, the model's shared component will learn the key characteristics of Portuguese speech, while the corpus-specific part will focus on how to apply them to adults and children, respectively.

In the fourth line, Training over children data with a pre-trained Portuguese adult model as initialization enhanced the result to 25.36% WER. When compared to weights random initialization, it is shown that the weights of the adult model are a beneficial starting configuration and allow the transfer learning model to learn relevant patterns for children. It avoids the need for the model to learn these patterns from scratch, using data from a highly variable source. 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 [8, 133].

3.2.5 Summary and discussion

In this study, we conducted a knowledge transfer technique analysis to improve the results of ASR systems for children. We corroborate the acoustic mismatch between adult and child speech and the importance of the model encountering child data and its variability. Our investigations revealed that the transfer learning 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, we focus on the transfer from adults to children. It is not clear how such a system can work using only children's data.

3.3 Multi-task and transfer learning using multilingual children data

3.3.1 Motivation

In this section, we study whether the performance of children's ASR for low-resourced languages may be improved by combining children's resources from different languages. In many cases, there is limited or no data for both adults and children. Therefore, we propose using several small-sized corpora of children from various languages to overcome the substantial acoustic variability and data scarcity issues. The current study extends standard multilingual training and transfer learning for hybrid HMM-DNN ASR by combining them in a meaningful way to use knowledge from heterogeneous data. First, a multilingual model trained with a multi-task learning objective tries to optimise network parameters to the specific characteristics of children's speech in multiple languages/tasks simultaneously. Subsequently, this multilingual model is used to improve ASR for a target language –potentially different from those used in the multilingual training stage– by using transfer learning. We address the following research question: Does this two-step training strategy outperform conventional single language/task training for 's speech, as well as multilingual and transfer learning alone?

3.3.2 Proposed approach

We propose to combine transfer learning (TL) and multi-task learning (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 multi-task learning objective attempts to optimize the network parameters to the particular characteristics of children's speech in multiple languages in parallel. In this work, the model is considered multilingual because all the tasks trained during multitask learning are a corpus of children from different languages. Secondly, we adapt this 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 are language-specific. That is, there are as many output layers as there are languages, i.e. children corpora. Notice 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 (figure 3.1). Our hypothesis is that the more data has been seen by the acoustic model, the better the shared layers can capture the underlying characteristics of children's speech during the first stage of the procedure. These characteristics can be

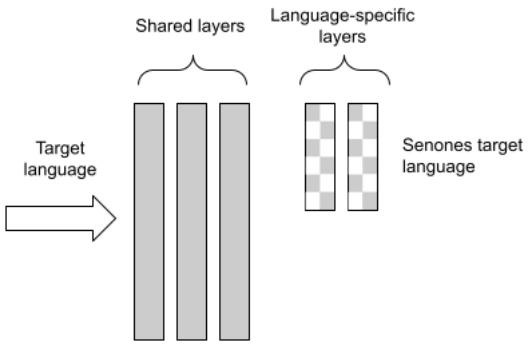


Figure 3.1: Multilingual transfer learning approach. Language-specific layers can be randomly initialized for a language not present during the MTL phase or use the corresponding pre-trained layers in case the target language was present during the MTL phase. Grey blocks are pre-trained during MTL phase.

used effectively, later, by the language-specific layers and during the second step of the procedure (figure 3.1).

Although the approaches adopted in this work have been used previously in other studies, for instance [133] and [140] 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 a methodological approach in the case of English cannot be expected to generalize to other contexts and languages. As we all know, English is a large-size, resource-rich pluricentric language which should be seen more as an exceptional case, rather than an average representative. Against this background, it is important to emphasize that there is a need for research that investigates whether methods that have already been tested for English also work in new contexts such as those of mid-sized languages with fewer resources than English, like Dutch, Portuguese, Swedish and German.

3.3.3 Setup

All experiments were conducted using five children corpora, each from a different language. Namely PFSTAR_SWE, ETLTDE, CMU, LETSREAD and CHOREC. All those datasets have been described in section 2.4.Table 5.1 presents statistics about the duration, number of utterances and language Notice that in this work we have only used small datasets to better reflect the average size of the available children's speech corpora.

We employ the same experimental design as the prior experiment with adult data from section 3.2.3 setup, where the acoustic model is divided in two. The first part is shared across all languages, whereas the second is language specific. Furthermore, each corpus, i.e. each language, uses an independent language model and lexicon that is constant throughout all experiments in order to assess solely the acoustic model contribution.

Corpus name	Language	Train	Test
PFSTAR_SWE	Swedish	6030 utt 04h00	2879 utt 01h48
ETLTDE	L2 German	1445 utt 04h41	339 utt 01h06
CMU	English	3637 utt 06h26	1543 utt 02h45
LETSREAD	Portuguese	3590 utt 12h00	1039 utt 02h30
CHOREC	Dutch	2490 utt 20h12	575 utt 04h42

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

3.3.4 Multilingual-transfer learning experiment

	PFSTAR_SWE	ETLTDE	CMU	LETSREAD	CHOREC
Language	*Swedish*	*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

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; Multi-task Training (MTL) alone, trained jointly using all corpora for 4 epochs ; Transfer Learning (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 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 [8].

Turning to multi-task learning, among all the approaches presented, only MTL fails to improve the baseline performance for almost all languages, which is in contradiction with [133]. However, it can be explained by the differences in terms of the size of the child's speech corpora used. The smaller the size of the corpora used, the more difficult it is to model the acoustic variation in the children's speech.

Concerning TL, all performance scores outperform their corresponding baseline, confirming that TL is an adequate method for children's ASR since it allows the system to be confronted with more children, thus with more variation. Precisely, table 3.4 shows that the best pre-trained model for knowledge transfer is Chorec. This makes sense since Chorec is the largest corpus, representing about 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) transfer learning 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, minimizing 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 optimization (lattice-free MMI) can partially explain these results. Nevertheless, we hypothesize that the multi-task objective leans the network towards better optimization of the lower layers, rather than optimizing 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.

3.3.5 Cross-lingual validation

In the previous section, we saw that the MLTL approach yields better results than separate multi-task and transfer-learning 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 MLTL 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 3.3.3 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.

3.3.6 Summary and discussion

In this work, we addressed the following research question: Does the two-step training strategy we propose in the current chapter outperform conventional single language/task training for children's speech, as well as multilingual and transfer learning alone. Our results provide a positive answer to this question, by showing that the limitations of MTL and TL can be overcome by the multilingual transfer learning approach, even in a low-resource scenario, leading to an average relative improvement of 7.73%. Multilingual pre-training is also beneficial for transfer learning with an unseen language, with an average relative improvement of 5.56%. Multilingual transfer learning thus seems to be an appropriate method to address children's speech recognition in a challenging context.

3.4 Conclusion

In this chapter, we look at the current state of the art for a Hybrid HMM-DNN speech recognition system for children. We illustrated that transfer learning is the most promising strategy for addressing children's ASR variability because it makes efficient use of the knowledge contained in the pre-trained

source model. A pre-trained model that can be trained on both adults or children. The multi-task learning does not produce the greatest results alone, but we showed that the shared part of the model is capable of learning relevant information for all tasks jointly. Furthermore, we proposed in this chapter to combine these two approaches in our multilingual transfer learning system. Using the capacity of learning relevant information of the multi-task learning approach and the capabilities of efficient use of pre-existing knowledge from transfer learning.

4

End-to-End children automatic speech recognition

Contents

4.1 Introduction

With the growing popularity of deep learning, numerous successful attempts to apply it to ASR have been made. It was only recently, that end-to-end models have shown their capability to outperform hybrid HMM-DNN systems for a variety of speech recognition tasks, including children's ASR. The major advantage of end-to-end speech recognition systems is the merging of the whole training process into a single neural network that eliminates the possibility of behavioural incompatibilities between modules that have been trained independently. However, because of the problem of children's data scarcity, the application of the end-to-end paradigm for children's ASR is relatively new and has not been extensively investigated [109–112]. In addition, end-to-end models often require more parameters to provide such robustness and flexibility. As a result, training on small datasets becomes increasingly difficult [181].

With the recent increased interest in end-to-end speech recognition, several architectures have been developed, including recurrent neural networks [182] and neural transducers [183]. However, one architecture stands out and consistently provides state-of-the-art results in large-vocabulary speech recognition for both adults and children, the Transformer.

This section dives more into details of the Transformer design as well as the adapter transfer for children ASR, a parameter-efficient transfer for Transformer models that we have recently proposed.

4.2 Transformer models

First proposed in 2017 [9], the Transformer architecture is a sequence-to-sequence encoder-decoder architecture that relies entirely on self-attention, eliminating recurrence, convolutions entirely and vanishing gradient issues. 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 encoding the symbol position in the sequence. This, enables more parallelisation, resulting in faster training. Since its release, the Transformer architecture had tremendous impact in various areas, including NLP [64], computer vision [184], and speech [82]. The transformer encoder-decoder architecture is presented in figure 4.1, with c) the encoder and d) the decoder. The encoder's role is to transform an input sequence $X = x_1, \ldots, x_T$ into a series of continuous representations $Z = z_1, \ldots, z_T$ which are then fed into a decoder. The decoder, constructs an output sequence $Y = y_1, \ldots, y_N$, one element at a time. At each time step, the decoder receives the encoder outputs together with the last decoder output, in an auto-regressive manner.

The information about the relative position of the tokens in the sequence is given by the summation between the input/output embedding and the positional embedding. Although there are many choices

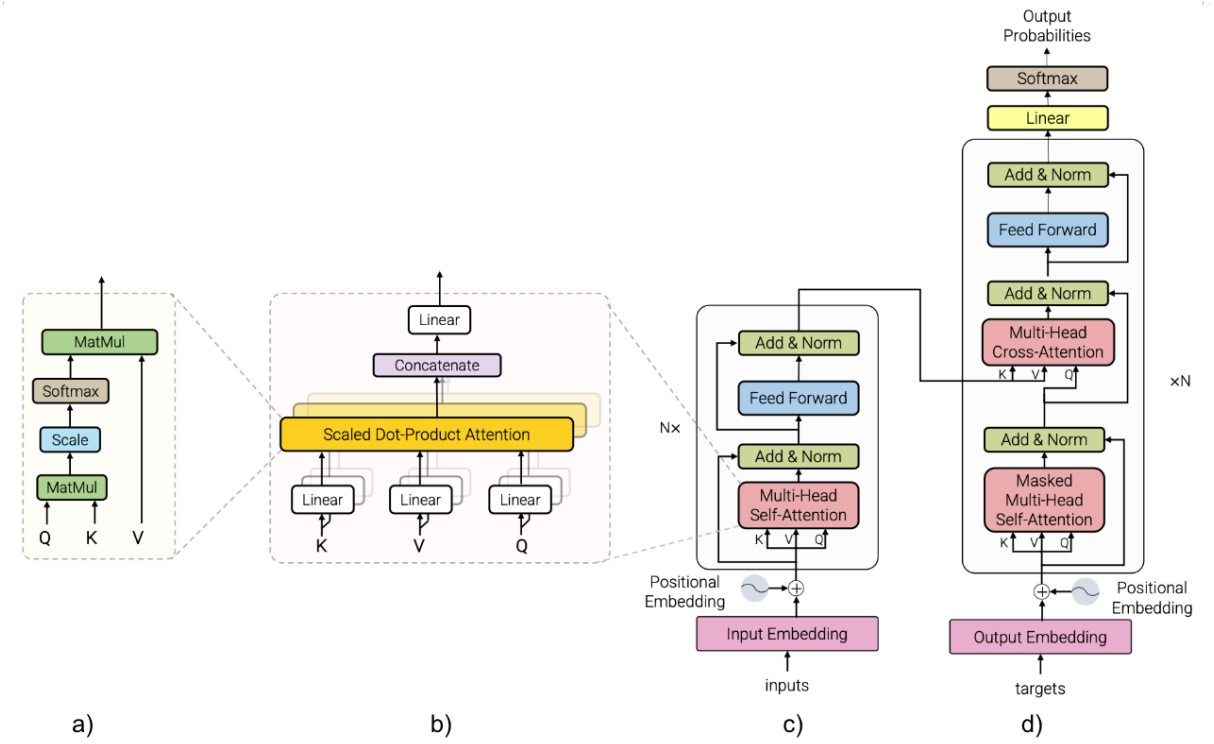


Figure 4.1: Architecture of the standard Transformer [9]. a) scaled dot-product attention, b) multi-head self-attention, c) Transformer-encoder, d) Transformer-decoder.

of positional encodings, [9] proposed to use sine and cosine of different frequencies as follows:

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

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

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

The encoder is composed of a stack of N identical Transformer layers. Each layer consists of a multi-head self-attention module and a feed-forward fully connected neural network module. Each of these modules are followed by a normalization with a residual connection. A multi-head self-attention (MHA) modules relies on the scaled dot-product attention [9], illustrated in figure 4.1.a). Scale dot-product attention focuses on how relevant a particular token is with respect to other tokens in the sequence. And is defined as follows:

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

Where 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 each divided by $\sqrt{d_k}$, then passes through a softmax function to

obtain attention weights on the values. When d_k is large, the scaling $\frac{1}{\sqrt{d_k}}$ restrains the dot product from growing large in magnitude.

Instead of performing a single scaled dot-product attention, the MHA module linearly projects h times K, V and Q with different, learned, linear projections to dimensions d_k,d_k and d_v respectively. On each of h projected versions is performed the attention function 4.3 in parallel. The output of each of the h attention function, of dimension d_v is concatenated and projected one last time as pictured in figure 4.1.b). More formally:

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \\ \text{where } head_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{4.4}$$

Where the different projections 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 $W^O \in \mathbb{R}^{hd_v \times d_{model}}$.

In comparison to the encoder, the decoder contains a third sub-layer, which performs MHA over the output of the encoder using the prior sub-layer of the decoder as the query. The decoder's inputs, which are targets during training and the previously decoded label during inference, are offset by one position. This combined with a modified MHA prevents the attention to use subsequent positions, ensuring that the prediction at time-step i solely depends on the previous $< i$ time-steps.

4.3 Adapters for Transformer based models

Age-dependent acoustic models have shown promising improvements, as children's age is highly correlated with acoustic variability [1, 31]. In particular, some studies found that variability decrease with the age, reaching the adult level at 15 years old [2]. In parallel, research on End-to-end (E2E) architectures has shown equivalent or even superior performance in a large number of speech recognition tasks compared to traditional hidden Markov models approaches [75]. E2E architectures propose to combine different modules of the ASR pipeline into a single deep neural network (DNN), resulting in benefits to avoid error accumulation and mismatch between components. However, for these models to work properly, they need to be trained with a large amount of data, which is not commonly available for children's speech. Thus, to overcome children's data sparsity issue for E2E models training, [109, 110] successfully used transfer learning by fine-tuning an adult pre-trained model on children's speech.

In this work, we propose to apply adapter modules on top of an adult acoustic model as an alternative to the transfer learning strategy for automatic children's speech recognition. Adapters are a method recently proposed for Transformer-based systems that consist of a small set of additional layers that are attached to a source model [10, 185]. Adapters are typically less expensive both in terms of training speed and storage cost, which is a desirable property in the case of aiming at the development of children's

age-dependent models. In addition, adapters overcome the problem of catastrophic forgetting. Indeed, after using transfer learning, the source model is completely overwritten by the newly trained weights, leading to a drop of performance on the source task. Whereas in adapter transfer, the backbone model remains frozen, thus preserved if adapter layers are removed. Adapters are therefore very practical in the context of small device computing where it can be expensive to load and store a large number of models for adults and children of different ages. Finally, in this work, we also propose a novel version of adapter layers inspired by variational auto-encoders (VAE) [186], so-called variational adapters or Vadapters. We hypothesize that the ability of VAEs to estimate variability can be applied in adapters to make them more suitable for parameter-efficient automatic children's speech recognition.

4.3.1 Related work

4.3.1.A Transformer model for children ASR

Recently, E2E-based ASR models have demonstrated their ability to achieve state-of-the-art performance on a wide variety of speech recognition tasks [75]. This fact motivated the assessment and comparison of different E2E architectures for children ASR [109, 110]. These works found that Transformer-based architectures, described in the previous section 4.2, yield the best results when an adult pre-trained model is fine-tuned for children using transfer learning with the help of the joint attention and CTC objectives [76]. Usually, these two objectives are combined as follows:

$$\mathcal{L}_{ASR} = \lambda_{ctc}\mathcal{L}_{ctc} + (1 - \lambda_{ctc})\mathcal{L}_{s2s} \tag{4.5}$$

where $\mathcal{L}_{ctc}$ and $\mathcal{L}_{s2s}$ are the CTC and attention losses, respectively. A hyper-parameter $\lambda_{ctc} \in [0, 1]$ is used to control contribution of each loss.

4.3.1.B Adapters

Adapters were first introduced in the NLP field, motivated by the need for a parameter-efficient adaptation to fine-tune large models, like Transformer, on various text classification tasks [185]. They are a simple alternative to full model fine-tuning, as they involve only a small number of newly inserted parameters at each layer of the transformer. While different positions have been proposed [10, 185], they are generally plugged after the feed-forward layer (see Figure 4.2.a). The key idea for training adapters is to freeze the backbone model's parameters and only update the adapter's parameters. Adapter modules are based on a bottleneck architecture (projection-down followed by a projection-up) as shown in Figure 4.2.b. Adapters solve a number of drawbacks associated with full model fine-tuning, such as parameter efficiency, faster learning iterations and a highly modular design

Since it was first proposed, adapters have been successfully used in a wide range of NLP tasks such

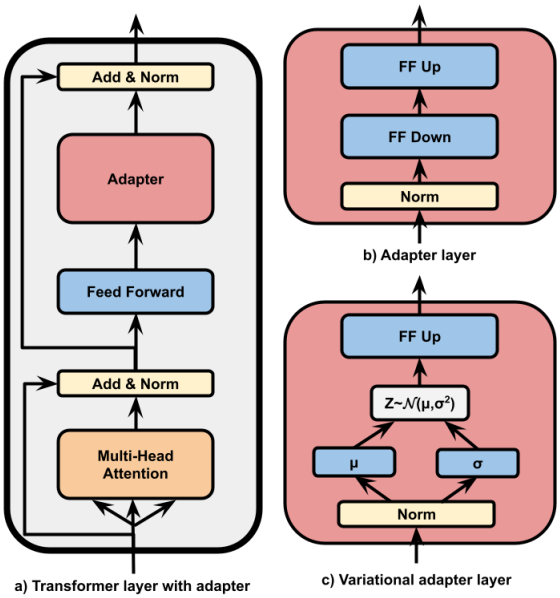


Figure 4.2: a) Example of a Transformer layer with an adapter layer (adapted from [10]); b) Adapter layer; c) Vadapter layer

as language understanding [10] and neural machine translation [187]. Some researchers proposed to use adapters for ASR tasks, such as in multilingual ASR [188]. More recently, [189] studied adapters for atypical speech, in particular pathological and accented speech. More recently, [190] proposed to use adapters inside of self-supervised models for children ASR by refining the whole model together with the weights of the adapters. Our work differs because our aim is to update only the adapters' weights in order to keep both the parameter efficiency and modular properties of adapters.

4.3.1.C Variational Auto-Encoders

Variational auto-encoders (VAE) [186] are a probabilistic generative models, that has been successfully applied in different speech tasks such as transformation [191] and enhancement [192]. The main strength of VAE is their ability to learn a smooth representation of the latent space. Indeed, rather than producing a single value to describe each element of the latent space, as a standard auto-encoder, VAE provides a probability distribution:

$$p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z}) \tag{4.6}$$

where $\mathbf{x}$ is the observed data generated by a random process using latent data $\mathbf{z}$ and θ denotes the distribution parameters. In this model, the likelihood function $p_\theta(\mathbf{x}|\mathbf{z})$ quantifies how the generation of $\mathbf{x}$ is conditioned by $\mathbf{z}$, while the prior $p_\theta(\mathbf{z})$ is used to regularize the latent data $\mathbf{z}$. Typically, a standard

Gaussian distribution is used for the prior distribution

$$p_\theta(\mathbf{z}) = \mathcal{N}(\mathbf{z}; 0, I) \tag{4.7}$$

while the likelihood is defined as a multivariate Gaussian distribution:

$$p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}; \mu_\theta(\mathbf{z}), \sigma^2_\theta(\mathbf{z})) \tag{4.8}$$

where $\mu_\theta(\mathbf{z})$ and $\sigma^2_\theta(\mathbf{z})$ are obtained using $\mathbf{z}$. However, since the posterior distribution $p_\theta(\mathbf{x}|\mathbf{z})$ is intractable, it is approximated with the auxiliary distribution $q_\phi(\mathbf{z}|\mathbf{x})$ that plays the role of an encoder:

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \tilde{\mu}_\phi(\mathbf{x}), \tilde{\sigma}^2_\phi(\mathbf{x})) \tag{4.9}$$

We also want to ensure that the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ and the true posterior $p_\theta(\mathbf{z}|\mathbf{x})$ are similar by minimizing the Kullback-Leibler (KL) divergence between the two distributions.

$$\min KL(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})) \tag{4.10}$$

It is possible to minimize expression (4.10) by maximizing the following expression as shown in [191]:

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \log p_\theta(\mathbf{x}|\mathbf{z}) - KL(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})) \tag{4.11}$$

Where the first term is the reconstruction error and the second term a regularisation.

Thus, the VAE loss function can be define as followed:

$$\mathcal{L}_{VAE} = \mathcal{L}_{recons} + \mathcal{L}_{KL} \tag{4.12}$$

$$= \mathcal{L}_{recons} + \sum_j KL((q^{(j)}_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})) \tag{4.13}$$

for each dimension j of the latent space.

4.3.2 Variational adapters

Adapters and auto-encoders (AE) share a similar encoder-decoder structure. Although the purpose of these two architectures is different, the role of their encoders is similar: map relevant characteristics of the input into a unique latent vector. On the other hand, their architecture differs in the decoders: AEs use the decoder to reconstruct the input, while adapters project the information contained in the latent vector to be processed by the next layer. Consequently, adapters suffer from the same problems as AEs, a poor capability to model variability. In order to be more robust to the high variability of children's speech,

we propose to represent each latent value in probabilistic terms. To this end, we propose Vadapter, a new adapter architecture in which the encoder structure of the adapter is replaced with the structure of a VAE's encoder as shown in Figure 4.2.c.

Consequently, during training, instead of a down-projection that maps the input into the latent representation, we now have two branches, producing the mean μ and variance σ. During inference, μ is used directly as a deterministic latent vector, discarding σ. We hypothesise that this deterministic inference allows Vadapters to capture variability in the σ branch while keeping the μ more robust. In addition, dropping the σ branch during inference keeps the number of parameters equivalent to normal adapters, thus preserving the parameter efficiency.

Similarly to VAEs, the regularisation term which ensures that the distribution of $q_i(\mathbf{z}|\mathbf{x})$ for each Vadapter at layer i is similar to the standard normal distribution $p(\mathbf{z})$ is required. However, as there are many Vadapter layers we normalise the sum of all regularization terms by the number of Vadapter layers:

$$\mathcal{L}_{KL_{all}} = \frac{\sum_L^i KL(q_i(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))}{L} \tag{4.14}$$

where L is the total number of Vadapters in the model. Then, we inject this regularisation loss into the E2E ASR loss defined in equation (4.5) as follows:

$$\mathcal{L}_{ASR} = \lambda_{ctc}\mathcal{L}_{ctc} + (1 - \lambda_{ctc})\mathcal{L}_{s2s} + \beta\mathcal{L}_{KL_{all}} \tag{4.15}$$

where β is an hyper-parameter to control the regularization's contribution.

4.3.3 Experiments

4.3.3.A Corpus

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

In this work, we decided to use the Boulder Learning My Science Tutor (MyST) corpus, described in section 2.4 given the task assigned to the children, which is to speak spontaneously. Indeed, the end-to-end model encapsulates the acoustic model and language model in the same network. As a result, if we train a model with a restricted amount of prompts on a data set of reading tasks, the model will learn and overfit the prompts. Thus, yielding unreliable results. Furthermore, for the purposes of our experiments, we decided to remove all utterances shorter than one second and longer than 20 seconds. .

The details of the filtered corpora used in our work are presented in Table 4.1.

4.3.3.B Implementation details

All experiments were performed using the SpeechBrain toolkit [193]. We used 12 Transformer layers for the encoder and 6 Transformer layers for the decoder, all with dimensions 512. This model has been pre-trained using the LibriSpeech dataset [33] and is publicly available[1]. Furthermore, for all of our experiments, we used the same Transformer language model, trained on 10 million words in order to only evaluate the contribution of the E2E model. The adapter architecture consists of a linear layer down-projection from dimension 512 to 256 with a ReLu activation, followed by a linear layer up-projection from dimension 256 back to 512. We analysed the influence of the latent dimension, i.e. the number of parameters, on the ASR performance in more detail in section 5.6. The Vadapters σ and μ branches consist of a linear layer down-projection from 512 to 256 each, while the up-projection remains the same as the aforementioned traditional adapters. For all the experiments, models were trained with a batch size of 16, $\lambda_{ctc} = 0.3$ and the same decoding hyper-parameters. All models were trained with a learning rate of 1.5e-4, for 40 epochs. Finally, for all our Vadapter experiments we choose $\beta = 1.0$.

4.3.3.C Experiments description

In our first experiment, we will attempt to determine which component of the transformer model is most important to ASR children. As a result, this information will be used to determine the best location of the adapters in the transformer layer. Indeed, the adapter should come after the most important component since it will project the output of that component into the expected transformer space. In order to do this, we studied the role of each transformer layer sub-module by fine-tuning one or two of them with the children's speech data.

Secondly, we analyze the performance of adapters in three scenarios: i) Adapters in all layers of the E2E model, ii) adapters only present in the encoder layers, and iii) adapters only in the decoder layers. These experiments are motivated by the fact that the encoder is closely related to the acoustics generating a high-level representation of speech, while the decoder generates output tokens related to the linguistic domain. The objective is then to evaluate which components, the encoder (acoustics) or the decoder (linguistics), benefit more from the adapter transfer. In order to compare our new architecture with traditional adapters, we reproduce the three scenarios mentioned above by replacing the adapters with our Vadapters. Furthermore, we evaluate the combination of Vadapter and traditional adapter in two scenarios, Vadaper in the encoder and adapter in the decoder, and vice versa.

[1] https://huggingface.co/speechbrain/asr-transformer-transformerlm-librispeech

4.3.4 Results

4.3.4.A Transfer learning experiments

Fine-tuned part	WER ↓	Trained parameters
None	25.04%	-
Full model	13.50%	71.5M
Norm	18.08%	57.9K
MHA	13.40%	25.2M
FFN	**12.57%**	37.8M
MHA + FFN	12.78%	63.0M
Norm + FFN	12.92%	37.9M
Norm + MHA	13.52%	25.3M

Table 4.2: Results of the fine-tuning on part of the model only

Table 4.2 shows results of the transfer learning on sub-modules of the Transformer model. Fine-tuning all the transformer's parameters, in the same way as the previous work [109,110], gives better results than using the model trained only on adult data with 13.76% compared to 25.04% WER respectively. The fine-tuning of all normalisation weights improved the score compared to the adult model with 18.04% but still under-perform compared to the full fine-tuning. Thus, the normalisation contribution in the children's transfer learning is limited. In contrast, fine-tuning the MHA or FFN yields better, results compared to the full transfer learning with 13.40% and 12.57% WER respectively. While always outperforming a full model update, the use of transfer learning on a combination of different model components reduces performance when compared to FFN alone. Transfer learning becomes more difficult by updating the weights of all components of the transformer as well as the non-transformer weights (i.e., Convolution blocks and embedding blocks), which explains why the entire fine-tuning produces worse results. In conclusion, FFN modules are the most relevant to fine-tune using transfer learning. This is because transformer feed-forward layers are key-value memories [194], where each key correlates with patterns in the training examples, and each value produces a distribution over the outputs. Consequently, adapters should be placed after the FFN sub-modules in order to achieve better results. This is consistent with Pfeiffer's work for NLP tasks [10].

4.3.5 Adapters and Vadapters results

4.3.5.A Adapters for children ASR

Table 5.2 presents the word error rate (WER) results of the different approaches. Firstly, the pre-trained Transformer adult model without any adaptation gives the worst result, with a WER of 25.04%, while adult performances on Librispeech corpus are usually less than 6%. This result shows the impact of the variability in child speech. Secondly, the adaptation of all 71.5 million parameters for children's speech

Method	WER	Trained parameters
No fine-tune	25.04%	-
Fine-tune	13.50%	71.5M
Adapter	14.33%	4.8M
Adapter encoder only	14.56%	3.2M
Adapter decoder only	20.10%	1.6M
Vadapter	14.19%	7.1M (4.8M)
Vadapter-enc + Adapter-dec	14.05%	6.3M (4.8M)
Adapter-enc + Vadapter-dec	14.35%	5.5M (4.8M)
Vadapter encoder only	14.51%	4.7M (3.2M)
Vadapter decoder only	20.23%	2.4M (1.6M)

Table 4.3: Results of the different approaches; In parenthesis are shown the number of parameters needed for inference after dropping the σ branch.

resulted in a considerable improvement, with 13.50% WER. This result correctly reflects the state-of-the-art performance obtained in the literature for the MyST corpus [110]. Regarding adapters, similarly to previous work in NLP [185] and ASR [189], we observe that they perform slightly worse than fine-tuning, with a score of 14.33% WER. However, it is important to note that adapters require less than 10% of all parameters of the full fine-tuning. We also investigate adapter transfer for encoder and decoder only. Table 5.2 shows that adapters are more relevant when plugged into the encoder with 14.56% WER while compared to the decoder with 20.10% WER. This result confirms that acoustic variability plays a critical role in the degradation of children ASR performance [195].

Additionally, we also evaluated how different adapter hidden-dimension, i.e. the number of parameters, influence the speech recognition performance compared to the fine-tuning model. Figure 4.3 displays the relative WER delta over the ratio of trainable parameters compared to the fine-tuned model. As a reference, the relative WER delta of the adult model with respect to fine-tuning is 85.5%. We observe that the performance difference between fine-tuning and adapters decreases as the number of trainable parameters increases. While with only 2% of trainable parameters, adapters manage to surpass by a large margin the source model performance, adapters need a minimum amount of parameters to get close to the fine-tuning performance. Nevertheless, adapter transfer outperforms full fine-tuning, when the number of parameters used is around 30% of the number of the entire model. There is therefore a trade-off between performance and parameter efficiency. A similar observation has been made in [190].

4.3.5.B Variational-adapters

Concerning the Vadapter architecture, with the exception of the cases where the Vadapters are placed in the decoder, all the scores are higher than their conventional counterpart, approaching the full fine-tuning score. The best configuration, Vadapter in the encoder and adapters in the decoder, reaches 14.05% WER. In the same way, as for the conventional adapters, we can see that the Vadapters are more advantageous when placed in the encoder since the score is 14.51% WER for Vadapters in the encoder

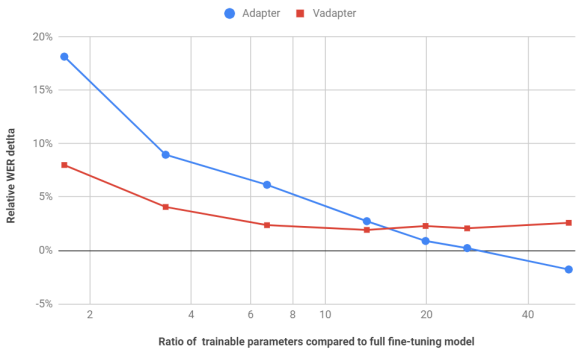


Figure 4.3: Relative WER delta over the ratio (%) of trainable parameters compared to full fine-tuned model.

only and 20.23% WER for the decoder only. We believe that this is because Vadapters are designed to be more robust to acoustic variability, which is mainly present at the encoder level. Thus, the Vadapters in the decoder does not manage to improve the score of their conventional counterpart.

Thus, we tested several possible combinations between Vadapter and adapters. We observe that the configurations with Vadapters in the encoder are giving the best results, with 14.19% WER for Vadapters in both the encoder and the decoder, as well as 14.05% for Vadapters in the encoder and adapters in the decoder. However, when Vadapters are placed in the decoder in combination with adapters in the encoder the result is not as good as adapters everywhere with 14.35% WER. We believe again that this is due to the variability being more present in the acoustic than in the linguistic component.

Finally, as shown in Figure 4.3, Vadapter outperforms conventional adapters when the number of parameters is less than 15% of the full model. Indeed, the Vadapters are always under 10% relative WER delta compared to full fine-tuning and reach under 5% with less than 4% of the ratio of trainable parameters, where conventional adapters start above 15% relative WER delta and need more than 10% of the ratio of trainable parameters to be under 5% relative WER delta. These results confirm the proposed Vadapter architecture as a more convenient alternative for parameter-efficient transfer learning. However, when the number of parameters increases, the results drop compared to conventional adapters. We hypothesize that this is due to the more complex and subject to variability sampling of $\mathbf{z}$ during Vadapters training.

4.3.6 Discusion

In this work, we demonstrate the usefulness of adapter transfer in the context of children's speech. With less than 10% of the total number of fine-tuning parameters, adapters are able to efficiently model children's speech. Noticeably, the adapter performance approaches fine-tuning, as the number of parameters increases. Furthermore, our Vadapter architecture outperforms conventional adapters in terms of acoustic variability robustness in a parameter-efficient setting. Using a combination of Vadapters in the encoder

and conventional adapters in the decoder allows for further improvement, getting closer to the fine-tuning performance while keeping a small number of parameters. This seems to demonstrate their effectiveness in modelling highly variable data, such as children's speech.

4.4 Summary

We covered the state-of-the-art for end-to-end children's speech recognition in this chapter. Particularly, the usage of the Transformer architecture in conjunction with transfer learning. In a similar way as chapter **??**, to avoid a drop in performances attributable to an acoustic mismatch between children and adults, the end-to-end model should be trained with children's data. In contrast to previous work, we demonstrated that fine-tuning only a portion of the transformer modules, particularly the FFN sub-module, yields better results since it serves as a key-value memory. As a result, we placed our adapter subsequent to it. The adapter's role is to accomplish knowledge transfer, which is related to transfer learning. Rather than updating the complete model's weights, we just tweak an extra module, hence fewer parameters. This adapter transfer achieves almost identical results as the entire model fine-tuning. In addition, adapters are useful in the context of customized models, where training and storing a whole model for each age group or each child can be expensive and time-consuming.

In addition, we proposed the variational adapter, a variant of the traditional adapter based on variational auto-encoders. Compared to the adapter, which takes a bottleneck encoder-decoder structure with a linear layer as encoder and a linear layer as decoder, the variational adapter's encoder consists of two branches, μ and σ. The outputs of these two branches are used as the mean and variance vector to sample the input of the decoder. By doing so, we enforce the adapter's input variability to be contained in the σ branch. A branch which is suppressed during inference. As a result, we reduce input variability while maintaining the same size as standard adapters.

5

Use of synthetic speech as data augmentation

Contents

Children's automatic speech recognition (ASR) poses a significant challenge due to the high variability nature of children's speech. The limited availability of training datasets hampers the effective modelling of this variability, which can be partially addressed using a text-to-speech (TTS) system for data augmentation. However, generated data may contain imperfections, potentially impacting performance. In this work, we use Adapters to handle the domain mismatch when fine-tuning with TTS data. This involves a two-step training process: training adapter layers with a frozen pre-trained model using synthetic data, then fine-tuning both adapters and the entire model with a mix of synthetic and real data, where only synthetic data passes through the adapters. Experimental results demonstrate up to 6% relative reduction in WER compared to the straightforward use of synthetic data, indicating the effectiveness of adapter-based architectures in learning from imperfect synthetic data.

5.1 Introduction

Advances in deep learning and large amounts of training data have greatly improved automatic speech recognition (ASR). ASR is now widely used in applications like automatic transcription and home assistants. However, the recognition of children's speech remains challenging. This is because children's speech differs significantly from adults due to developmental changes in their speech production apparatus [**?**]. These changes lead to shifts in fundamental frequencies, altered temporal and spectral characteristics, and increased disfluencies [2, 27]. Additionally, children's limited linguistic and phonetic knowledge further complicates speech recognition [104]. Finally, the scarcity of children's speech data also limits the robustness of ASR models to these variabilities.

To bridge the gap between adult and children's automatic speech recognition, significant improvements have been made to the ASR pipeline. These advancements include techniques like Vocal Tract Length Normalization (VTLN) [90], pitch and formant modification [96], and adversarial multi-task learning [100]. In [**?**], a large dataset of children's speech, comparable in size to an adult corpus, was used to train an ASR model. This system achieved state-of-the-art performance, showcasing that neural networks can effectively learn from diverse and variable children's speech data when there is ample training data. However, acquiring and annotating datasets for training speech recognition models can be notably challenging and costly, especially for children, as their attention span is limited, and they may not consistently follow prompts during reading tasks. An alternative approach is to generate synthetic datasets using a text-to-speech (TTS) model. TTS can bypass the difficulties of collecting and annotating real children's speech data. Some studies have explored TTS for ASR, either by directly using synthetic speech for training or as a form of data augmentation [120]. However, synthesizing children's speech is challenging due to their inherent substandard and imprecise pronunciation [121]. Therefore, using synthetic data directly could lead to a performance decrease [121, 196].

In this paper, we introduce a novel technique called "Adapter double-way fine-tuning" to enhance

ASR models for children, even when using imperfect data augmentation. Our approach involves adding additional adapter layers to the existing ASR model during fine-tuning, similar to [190]. These adapter layers are customised to address the domain mismatch between real and synthetic data. We achieve this through a two-step training procedure. In the first step, the adapter layers are trained exclusively using synthetic data while keeping the pre-trained model frozen. In the second step, we fine-tune both the trained adapters and the entire model using a combination of synthetic and real data. Crucially, our approach differentiates between synthetic and real data during fine-tuning. Synthetic data passes through the adapter layers, while real data bypasses them. This approach enables the effective use of imperfect synthetic data to enhance ASR performance for children.

5.2 Related work

5.2.1 TTS data augmentation

The advancement of TTS systems, achieving human-like quality, enables effective TTS-based data augmentation in ASR. This approach, as shown in studies like [120], involves generating synthetic speech from text using TTS models, then combining it with real speech for training, resulting in performance enhancements. Notably, this approach is not limited to well-resourced tasks and has succeeded in low-resource scenarios, as demonstrated in [**?**]. Nevertheless, TTS data augmentation offers only modest improvement due to the domain mismatch between synthetic and real speech. In order to mitigate the mismatch with real data and to reduce speaker dependency, the use of discrete representations based on VQ-wav2vec has been proposed [**?**].

Another approach to mitigate the domain mismatch between synthetic and real speech is through the use of data selection techniques, as suggested by [121]. By selectively choosing high-quality synthetic speech data. This data selection process ensures that only the most reliable and accurate synthetic speech samples are used during data augmentation. The results presented in [121] demonstrate the effectiveness of employing i-vector speaker-embedding cosine similarity as a metric for data selection, compared to metrics like error rate, acoustic posterior, and synthetic discriminator.

In Synth++ [196], an extension to the data selection technique is proposed, by incorporating separate batch normalization statistics for real and synthetic samples. While data selection handles artefacts and over/under-sampling, double batch normalization aims to further bridge the synthetic-real data gap during training. This approach uses rejection sampling based on a DNN's output. Where the DNN is trained on a 5-dimensional features vector derived from a pre-trained ASR model, including cross-entropy loss, CTC loss [**?**], word error rate (WER), lengths of tokens in prediction text, and length of tokens in target text, offering valuable insights into speech quality and characteristics.

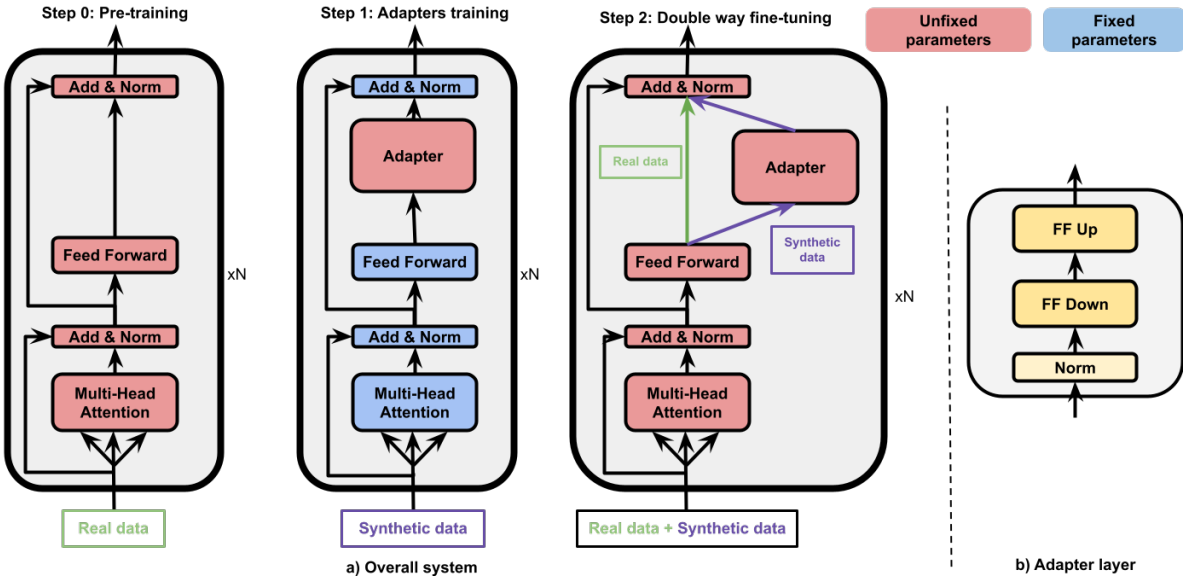


Figure 5.1: Overview of a) double way fine-tuning and b) Adapter layer architecture

5.2.2 Adapters

Adapters were first introduced for natural language processing (NLP) tasks as a simpler alternative to full model fine-tuning [185]. They involve adding a small number of extra parameters to each layer of the transformer model. Unlike full fine-tuning, which modifies the entire model, adapters enable targeted adjustments within specific layers while keeping pre-trained parameters intact. These adapters typically follow a bottleneck architecture with down-projection and up-projection, as seen in Figure 5.1-b. The bottleneck architecture's purpose is to introduce non-linear transformations, enabling Adapters to capture task-specific features and learn task-specific modifications effectively.

Since their proposal, adapters have demonstrated effectiveness in diverse NLP tasks, including language understanding and neural machine translation [187]. Additionally, there is a growing interest in applying adapters to automatic speech recognition. For instance, [189] explored adapters for atypical speech, focusing on pathological and accented speech. In children's ASR, Adapters are employed within the Draft framework [190]. This approach inserts and trains Adapters at each block of a pre-trained self-supervised learning (SSL) model using an SSL loss. Subsequently, the entire model, including the Adapters, undergoes fine-tuning with ASR losses. By combining SSL pre-training, Adapters, and full fine-tuning, this approach uses the advantages of SSL, the adaptability of adapters, and task-specific fine-tuning to enhance the recognition accuracy of ASR systems for children's speech.

5.3 Method

Expanding on the achievements of the Draft Framework and Synth++, our approach utilizes Adapters as a substitute for the double batch normalization layer of the Synth++ framework. Our aim is to improve the performance of a pre-trained ASR model through data augmentation using synthetic data. In our methodology, we employed filtered synthetic data, implementing a speaker-embedding cosine similarity metric to retain synthetic utterances that exhibited high-quality generation. Our approach introduces two extra steps following the standard ASR model training (Step 0).Figure 5.1-a provides an overview of our proposed methodology.

Step 1 entails training Adapter layers while keeping the ASR model parameters fixed. These Adapter layers are placed after the transformer layers' feed-forward component, aiming to learn a projection that aligns synthetic children's speech with real children's speech within the transformer layers. This step is crucial as Adapter layers require this learning process. Without it, the subsequent fine-tuning in Step 2 could be more challenging and less effective.

In Step 2, we fine-tune both the adapters from Step 1 and the pre-trained ASR model using a mix of synthetic and real data. A crucial aspect of our approach is how we handle data flow within the model. Real samples bypass the adapter layers as they don't need further adjustments, directly passing through the original ASR model components. Synthetic data, on the other hand, goes through the adapter layers for necessary modifications to align better with real children's speech characteristics. This differential treatment of data optimises adapter usage, potentially improving the ASR system's overall performance.

During inference, the Adapter layers become unnecessary and are discarded because the test data only contains real samples. It is important to mention that Steps 1 and 2 can be iteratively repeated with newly generated synthetic data, although this aspect is not investigated in this paper and is a subject for future research.

In summary, our approach uses adapter modules to improve the performance of a pre-trained ASR model through the integration of filtered synthetic data augmentation.

5.4 System description

5.4.1 Transformer architecture for ASR

The Transformer architecture, initially developed for tasks like machine translation [197], was found to be highly effective and widely used in various domains, including computer vision [**?**] and language understanding [64]. In speech recognition, it takes acoustic features as input, processes them through an encoder to create high-level representations, and uses these for token prediction in a decoder. Training typically combines a sequence-to-sequence approach with a CTC loss [**?**]. Recent studies, such as [110],

demonstrate that fine-tuning adult pre-trained Transformer-based models with children's speech data yield better results than traditional HMM-DNN based models, making the Transformer-based and End-to-end models a suitable choice for children's ASR.

In our experiments, we employed the SpeechBrain toolkit [193] for the ASR part of our system, using a pre-trained Transformer model[1], trained on the LibriSpeech dataset [33]. This model includes 12 encoder layers and 6 decoder layers, each having a dimension of 512. Additionally, we incorporated a Transformer language model trained on a 10 million-word corpus.

5.4.2 Multi-speaker text-to-speech: YourTTS

In this work, we used the pre-trained YourTTS[2] model proposed by [**?**] based on the Coqui toolkit. YourTTS is a TTS model that is built upon the Variational Inference with adversarial learning for end-to-end Text-to-Speech (VITS). It incorporates several novel modifications to enable zero-shot multi-speaker and multilingual synthesis. YourTTS use a 10-layer Transformer-based text encoder with 196 hidden channels. It can be used in a multilingual fashion by using a 4-dimensional language embedding concatenated with the embedding of each input character, for the purpose of our experiment, the multilingual aspect was discarded to only keep the English language. The decoder has four affine coupling layers, each with four WaveNet blocks for high-quality speech generation. YourTTS uses an external H/ASP speaker encoder to generate 512-dimensional speaker embeddings for individual speakers, serving as reference speakers for the model. Additionally, YourTTS incorporates a HifiGAN vocoder [**?**]. As YourTTS is an end-to-end model, the vocoder is connected to the TTS model using a variational autoencoder (VAE). For a comprehensive understanding of the YourTTS architecture and training, detailed information can be found in the original paper [**?**].

5.5 Experimental setup

5.5.1 Real speech corpus

Table 5.1: My Science Tutor Children Speech Corpus statistics

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

In this study, we used the My Science Tutor (MyST) Children Speech Corpus, referred to as the

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

"Real" set. This corpus contains around 400 hours of speech collected from 1,372 students in grades three to five. It comprises conversations with a virtual tutor spanning eight scientific domains. Notably, only 45% of the utterances in the corpus are transcribed. For our experiments, we filtered out utterances shorter than one second and longer than 30 seconds due to GPU memory constraints. Additional details on the filtered corpora are provided in Table 5.1.

5.5.2 Synthetic data

To adapt YourTTS for generating children's speech, we fine-tuned YourTTS using the MyST training set. In this study, we developed two TTS systems with different parameter settings to investigate their performance and output quality under varying conditions.

The first model, referred to as TTS_1, underwent fine-tuning for 250 epochs without including the speaker encoder loss. In contrast, the second system, labelled TTS_2, was fine-tuned for 50 epochs while incorporating the speaker encoder loss. This incorporation improved the alignment between the generated speech and the reference speaker embedding provided to the model.

The first TTS model, TTS_1, was used to generate 300 hours of synthetic data referred to as $Synth_1$. The second TTS model, TTS_2, was employed to generate a larger volume of synthetic data, up to 1,000 hours, denoted as $Synth_2$. To compare the performance of $Synth_1$ and $Synth_2$, a subset of 300 hours was extracted from the $Synth_2$ dataset. The full 1,000-hour set was exclusively used to evaluate the impact of different amounts of synthetic data, both reduced and increased.

In both $Synth_1$ and $Synth_2$, we used randomly selected d-vectors and text transcriptions from the MyST training set. Notably, the selected d-vectors did not match the associated transcriptions to introduce variability into the synthetic data. We used MyST transcriptions to generate synthetic data, exposing the TTS model to its unique transcription style, including elements like "UM" hesitations. This approach helps the model learn and reproduce the specific transcription characteristics of the MyST data. To assess the filtering effect, we generated an extra 300-hour set for both $Synth_1$ and $Synth_2$ without using speaker embedding data selection. Our data selection method relied on cosine similarity using x-vectors from a pre-trained x-vector extractor[3]. We applied a cosine similarity threshold of 0.75 to discard bad synthetic utterances. We also explored the data selection mechanism suggested by [196] but found it unsatisfactory, opting instead for speaker-embedding similarity as the selection criterion.

5.5.3 Experiments

We evaluated our Adapter double-way fine-tuning approach in a series of experiments, comparing it to existing methods. We started with baseline models fine-tuning an adult model to children's speech using real data for 20 and 25 epochs (step 0 in Figure 5.1-a). Next, we assessed the TTS models'

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

Method	$Synth_1$	$Synth_2$
Real (20 epochs)	12.99%	
Real (25 epochs)	13.15%	
Real + Non filtered *Synth*	13.41%	13.24%
Real + *Synth* [121]	13.09%	12.98%
Synth alone	40.58%	40.21%
Norm double-way (from adult)	12.89%	13.04%
Norm double-way (from children)	13.56%	13.87%
Adapter double-way (Ours)	**12.42%**	**12.31%**

Table 5.2: Results of the different approaches (in WER).

performances using only $Synth_1$ and $Synth_2$ data. We also explored data filtering's impact by comparing models trained on filtered and unfiltered versions of $Synth_1$ and $Synth_2$, along with their combination with *Real* data. These models were trained for 20 epochs. We also explored double-way normalization inspired by Synt++. In one scenario, we fine-tuned the adult model for 20 epochs using a mix of filtered synthetic and real data with double-way normalization (*Norm double-way from adult* in Table 5.2). In another scenario, we trained the double-way normalization model for 5 epochs with the baseline model as initialization, referred to as *Norm double-way from children*. Finally, we implemented our *Adapter double-way* approach, training the models for 5 epochs with the baseline model as initialisation. Different hyper-parameter configurations will be explored in section 5.6.

5.6 Results and discussion

5.6.1 Comparison with existing approaches

The results of the various approaches are summarized in Table 5.2. Our baseline models achieved a WER score of 12.99%. Training for 25 epochs led to over-fitting and a decrease in performance. Filtered $Synth_1$ and $Synth_2$ data improved WER by 2% compared to unfiltered data, but using only filtered TTS speech (*Synth* alone) resulted in a significant 40% WER on the *Real* test set, highlighting the domain mismatch between real and synthetic. Our experiments found that double batch normalization did not improve the baseline model's performance and even led to a 5% relative decrease in WER performance when evaluated with the baseline model as initialisation. This highlights the need for alternative methods to address the domain mismatch between real and synthetic speech data. Our double-way adapter fine-tuning approach, initialised with the baseline model (step 0 in Figure 5.1), outperformed all other methods. It achieved a 4% and 5% relative WER improvement over the baseline on $Synth_1$ and $Synth_2$ respectively, demonstrating the effectiveness of our approach compared to longer training on the *Real* set.

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 5.3: Results of the different number of hours in our Adapter double-way approach with $Synth_2$ data

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

5.6.2 Effect of the number of hours

Table 5.3 summarizes the impact of varying amounts of synthetic data from $Synth_2$ on our adapter double-way approach. Using a small amount of synthesized speech (10 to 50 hours) yields limited ASR performance improvement. While excessive TTS data (600 to 1,000 hours) can introduce noise. Thus, it's crucial to use an appropriate amount (100 to 300 hours) to balance between robustness and avoiding noise introduction.

5.6.3 Effect of the Adapters hyper-parameters

To assess the robustness of our approach, we assessed the Double-way adapter in diverse configurations. This involved experimenting with different bottleneck sizes (ranging from 64 to 2048), varying the number of training epochs (5 and 20), exploring the use of Adapters in the decoder of the transformer model, and conducting an ablation study by skipping step 0 and step 0 and 1.

Table 5.4 summarizes the results, highlighting that the optimal configuration uses adapters with a size of 1024 in the encoder only, coupled with 20 training epochs, resulting in a 6% relative WER reduction when compared to the baseline. Importantly, all configurations demonstrated superior performance to the baseline, underscoring the effectiveness of our approach.

Our findings suggest that extended training periods were beneficial for larger Adapter bottleneck sizes, without indications of overfitting. Moreover, adding Adapters to the decoder did not significantly improve results. Finally, skipping step 0 (pre-training) did not significantly degrade results, but skipping

both step 0 and step 1 (pre-training and Adapter pre-training) led to performance degradation, indicating the importance of Adapter pre-training for improved performance.

5.7 Conclusions and future work

We introduced the combined use of Adapters and synthetic data augmentation for children's speech recognition. Our two-step training procedure, involving training Adapter layers using synthetic data and subsequent fine-tuning of Adapters and the entire model with a combination of synthetic and real data, yielded improvements over the baseline and previous approaches in various configurations. or future work, we will explore an iterative approach with newly generated TTS data and varying the amount of real data used.

5.8 Ongoing and future work

In order to answer the following research questions: *Is it possible to develop an age-based, parameter-efficient automatic speech recognition model?; 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's variability?; - Given that self-supervised representation based ASR for adults matches or surpasses current state-of-the-art, are these representation appropriate for children's speech?* , we want to pursue three research directions in the future work of this thesis: As a first direction, we want to keep exploring adapter transfer. For example, as proposed by Pfeiffer [198], employing multiple adapters trained on different age groups or children corpus and combining them with an attention mechanism. It may also be interesting to investigate the use of explicit speaker information for robust adapter transfer [199]. Furthermore, as explained in section 4.3.2, Adapter module is structurally similar to an auto-encoder, thus it would be interesting to modify the structure of the adapter to follow the latest development in auto-encoder research. For instance, using neural discrete representations with the help of vector-quantized codebook [200] at the end of the adapter's down-projection.

As a second direction we want to investigate on the use of Text-to-speech (TTS) data augmentation. Indeed, as mentioned in section 2.1.3, one of the most significant obstacles in children automatic speech recognition is the lack of training data. One solution to this problem is voice conversion, in which the adult speech is transformed to child speech and then used for augmentation. However, the modified adult data can only capture a subset of the aspects of children's speech. As a result, it is important to generate children voice data directly from text [121]. A multi-speaker TTS system using speaker embeddings and text as input can be used to produce a synthesized utterance with the variability of child speaker. Indeed, the speaker embedding includes acoustic variability informations, which we want to find in the output utterance [119, 201]. One of the most challenging aspects of this approach is that the TTS model for

children produces unequal quality speech due mainly to acoustic variability, and hence the ASR system suffers when trained with this additional synthetic data. Furthermore, because we intend to augment the original data with TTS data, there may be a domain shift that worsens the ASR performance. To address both of these concerns, a speaker embedding-based data selection method has been suggested, based on the computation of the cosine-similarity between the input speaker embedding and the speaker embedding obtained from the output utterance. [121]. More recently, in order to minimise domain shift, two separate normalisation layers have been employed, one for the original data and one for the TTS data. [196]. In our research, we want to combine these two approaches to maximise the contribution of TTS data during training. In addition, we want to investigate whether the use of a GAN [202], which could create an artificial children embedding, could alleviate the problem of the limited number of speakers during training.

The final research direction we want to explore in this thesis is self-supervised learning (SSL) as a front-end feature rather than typical filter banks or MFCCs. For these models, the training process is separated into two stages. The first phase of training is self-supervised, which implies that no labels are used during training. The objective of this first phase is to present a large amount of unlabelled data to the system so that it learns a good speech representation. The second stage of learning is supervised fine-tuning, in which the model is taught to predict specific phonemes using the robust representation acquired in the previous stage with the help of a small amount of labelled data. In this category, two models stand out as state-of-the-art: Wav2Vec 2.0 [203] and HuBert [204]. As a preliminary experiment, to asses the usability of such frameworks for children ASR, we trained a BiLSTM model using the output of a variety of frozen self-supervised systems. For this experiment we used a subset of 50h of the Myst corpus [35], and the preliminary findings are displayed in the table 5.5

Front-end	UER	WER
Fbanks	12.29%	35.14%
TERA [205]	11.31%	31.80%
Audio Albert [206]	12.28%	34.69%
Wav2Vec2.0 Base	7.37%	19.76%
Wav2Vec2.0 Large	7.00%	18.76%
Distill HuBert [207]	9.22%	25.75%
HuBert Base	7.40%	19.77%
HuBert Large	**6.03%**	**15.41%**

Table 5.5: Results without language model of Self-supervised front-end

Where Base, and Large represent the same model with different number of parameters (in the order Base $<$ Large). Even though we did not use a language model in this pilot experiment, the results are of the same order as those reported in section 5.6 obtained with a transformer and a transformer language model. Such results demonstrate that SSL learns substantial speech characteristics. For future research, we aim to explore in depth what information is encoded in SSL models and why they work well on

children, and how we may use this knowledge to enhance children's ASR.

6

Pathology detection from speech

As mentioned in the previous section, SLT can assist paediatric speech therapists by automatically assessing pronunciation quality and identifying pathological conditions. Although the primary aim of this thesis was to improve ASR for reliable assessment of pronunciation quality. We also contributed to the identification of pathological conditions from speech, which will be discussed in this section.

The potential of speech as a non-invasive biomarker for evaluating a speaker's health for both physical and psychological disorders has repeatedly been proven by the results of several works [208, 209]. Traditional speech-based disease classification systems have focused on carefully researched, knowledge-based features. However, these features do not always capture the full disease's symptomatology and may even ignore some of its more subtle signs. This has led research to move towards generic representations that intrinsically model the symptoms. However, there are not enough pathological speech data available to train a large model directly. In our work [210], we proposed to assess speaker embedding, such as *i-vectors* [102] and *x-vectors* [211], applicability as a generic feature extraction method to the detection of Parkinson's disease (PD) and Obstructive Sleep Apnea (OSA). All disease classifications were performed with a support-vector-machine (SVM) classifier. Our experiments with European Portuguese datasets support the hypothesis that discriminative speaker embeddings contain information relevant to disease detection. In particular, we found evidence that these embeddings contain information that hand-crafted features fail to represent, thus proving the validity of our approach. It was also observed that x-vectors are more suitable than i-vectors for tasks whose domain does not match the training data, such as verbal task mismatch and cross-lingual. This indicates that x-vectors embeddings are a strong contender in the replacement of knowledge-based feature sets for PD and OSA detection.

Later, in [212], we proposed to extend the aforementioned work by classifying Alzheimer's disease with the conjunction of both acoustic and textual feature embeddings. In this end, speech signals are encoded into *x-vector* using pre-trained models. For textual input, contextual embedding vectors are first extracted using an English Bert model [64] and then used to feed a bidirectional recurrent neural network with attention. This multi-model system, based on the combination of linguistic and acoustic information, attained a classification accuracy of 81.25%. Results have shown the importance of linguistic features in the classification of Alzheimer's disease, which outperforms the acoustic ones in terms of accuracy.

Finally, we further extend the idea of using pre-trained representation to automatically detect COVID-19 from cough recordings. We leverage transfer learning to develop a set of COVID-19 classification subsystems based on deep cough representation extractors called experts. Individual decisions of three experts are fed to a calibrated decision-level fusion system. This ensemble of expert subsystems based on cough representations is expected to produce well-calibrated log-likelihood scores over a wide range of operating points. The output can be more easily interpreted by a human expert and incorporated into the decision-making process. Our results show competitive performance compared to hand-crafted features, although they are still far from those required to become a reliable tool to assist COVID-19

screening.

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] 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.

[7] 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.

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

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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

[24] 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.

[25] 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.

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

[27] 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.

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

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

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

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

[41] 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.

[42] 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.

[43] 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.

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

[45] 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

[46] 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.

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

[48] 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.

[49] 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.

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

[51] 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.

[52] 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.

[53] 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.

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

[55] 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.

[56] 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.

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

[58] 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.

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

[60] 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.

[61] 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

[62] 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

[63] 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.

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

[65] 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.

[66] 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.

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

[68] 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.

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

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

[71] 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.

[72] 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.

[73] 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.

[74] 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.

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

[76] 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

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

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

[79] 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.

[80] O. Vinyals and Q. Le, "A neural conversational model," *arXiv preprint arXiv:1506.05869*, 2015.

[81] 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.

[82] 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.

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

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

[85] 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.

[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. 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.

[88] 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.

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

[90] 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.

[91] Y. Saito, S. Takamichi, and H. Saruwatari, "Vocoder-free text-to-speech synthesis incorporating generative adversarial networks using low-/multi-frequency stft amplitude spectra," *Computer Speech & Language*, vol. 58, 06 2019.

[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] 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.

[94] 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.

[95] 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.

[96] 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.

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

[98] 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.

[99] 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.

[100] 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.

[101] 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.

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

[103] 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.

[104] 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.

[105] 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

[106] 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.

[107] 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.

[108] 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.

[109] 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.

[110] 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.

[111] 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.

[112] 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.

[113] 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 *Proc. Interspeech 2021*, 2021, pp. 1294–1298.

[114] 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.

[115] 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.

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

[117] 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.

[118] 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.

[119] 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.

[120] 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.

[121] 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.

[122] 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.

[123] 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

[124] 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.

[125] 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.

[126] 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.

[127] 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.

[128] 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.

[129] 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.

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

[131] 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.

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

[133] 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.

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

[135] 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

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

[137] 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.

[138] 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.

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

[140] 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.

[141] 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.

[142] 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.

[143] 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.

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

[145] 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.

[146] 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.

[147] 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.

[148] 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.

[149] 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.

[150] 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.

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

[152] 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.

[153] 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.

[154] 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.

[155] 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.

[156] 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.

[157] 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.

[158] 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.

[159] 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.

[160] 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

[161] 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.

[162] 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.

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

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

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

[166] 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.

[167] 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

[168] 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.

[169] 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.

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

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

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

[173] 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.

[174] 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.

[175] 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.

[176] 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

[177] 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.

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

[179] 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.

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

[181] 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.

[182] 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.

[183] 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.

[184] 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.

[185] 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.

[186] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *CoRR*, vol. abs/1312.6114, 2014.

[187] 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.

[188] 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.

[189] 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.

[190] 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.

[191] M. Blaauw and J. Bonada, "Modeling and transforming speech using variational autoencoders," *Morgan N, editor. Interspeech 2016; 2016 Sep 8-12; San Francisco, CA.[place unknown]: ISCA; 2016. p. 1770-4.*, 2016.

[192] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2018, pp. 1–6.

[193] 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.

[194] M. Geva, R. Schuster, J. Berant, and O. Levy, "Transformer feed-forward layers are key-value memories," *arXiv preprint arXiv:2012.14913*, 2020.

[195] 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.

[196] 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.

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

[198] J. Pfeiffer, A. Kamath, A. Rücklé, K. Cho, and I. Gurevych, "Adapterfusion: Non-destructive task composition for transfer learning," *arXiv preprint arXiv:2005.00247*, 2020.

[199] X. Gong, Y. Lu, Z. Zhou, and Y. Qian, "Layer-wise fast adaptation for end-to-end multi-accent speech recognition," *arXiv preprint arXiv:2204.09883*, 2022.

[200] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[201] E. Cooper, C.-I. Lai, Y. Yasuda, F. Fang, X. Wang, N. Chen, and J. Yamagishi, "Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6184–6188.

[202] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[203] 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.

[204] 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.

[205] 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.

[206] 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," 2020. [Online]. Available: https://arxiv.org/abs/2005.08575

[207] H.-J. Chang, S. wen Yang, and H. yi Lee, "Distilhubert: Speech representation learning by layer-wise distillation of hidden-unit bert," 2022.

[208] 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.

[209] 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.

[210] C. Botelho, F. Teixeira, T. Rolland, A. Abad, and I. Trancoso, "Pathological speech detection using x-vector embeddings," *arXiv preprint arXiv:2003.00864*, 2020.

[211] 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.

[212] A. Pompili, T. Rolland, and A. Abad, "The inesc-id multi-modal system for the adress 2020 challenge," *arXiv preprint arXiv:2005.14646*, 2020.