

# Master Thesis

Mixture-of-Experts for Age-Inclusive ASR: Reducing the Adult-Child Performance Gap in Multilingual Speech Recognition

Amelia Katarzyna Sasin

Delft University of Technology



# Master Thesis

## Mixture-of-Experts for Age-Inclusive ASR: Reducing the Adult-Child Performance Gap in Multilingual Speech Recognition

by

Amelia Katarzyna Sasin

In partial fulfilment of the requirements for the degree of  
**Master of Science**  
at Delft University of Technology,  
to be defended publicly on 20 July 2026

Faculty: **Faculty of Electrical Engineering, Mathematics and Computer Science**  
Department: **Intelligent Systems**  
Programme: **Data Science and Artificial Intelligence Technology**

Supervisors: Prof. dr. Odette Scharenborg (TU Delft)  
Dr. Piotr Żelasko (NVIDIA)

Thesis committee: Prof. dr. Odette Scharenborg (Chair)  
Dr. Bernd Dudzik (TU Delft committee member)  
Dr. Piotr Żelasko (External committee member)

An electronic version of this thesis will be available at [repository.tudelft.nl](https://repository.tudelft.nl).



# Preface

*This thesis concludes my MSc in Data Science and AI Technology at TU Delft, and the nine months of work behind it. My supervisor, Prof. dr. Odette Scharenborg, guided me throughout this project and gave me the opportunity to co-write a successful NVIDIA Academic Research Grant proposal with her; the compute it provided made this work possible. I did not expect, when we wrote that proposal, how significant this project would become for me: it has also led to my acceptance to the Young Female Researchers in Speech Workshop (YFRSW), held alongside Interspeech 2026 in Sydney, which I am very much looking forward to attending.*

*Dr. Piotr Żelasko provided daily supervision and technical expertise throughout this project, for which I am grateful, and I thank Dr. Bernd Dudzik for being part of my committee. I am grateful to my colleagues for supporting me and collaborating with me throughout these two years of my master's degree. I leave TU Delft having learned a lot, and I will miss it.*

*Amelia Katarzyna Sasin  
Delft, July 2026*

# Abstract

Automatic speech recognition (ASR) remains less accurate for children than for adults. This difference is poorly understood in multilingual systems. This thesis examines whether adding sparse Mixture-of-Experts (MoE) layers to a shared FastConformer-CTC encoder can narrow the adult-child word error rate (WER) gap. It also tests whether experts specialise by age, even without age labels in the training data. Dense and MoE encoders were initialised from the same pretrained checkpoint and fine-tuned in two stages: first on adult speech, then on mixed adult-child speech. Experiments covered English, Dutch, German, and Polish. Paired speaker-level bootstrap tests compared WER gaps. In the English-only experiment, MoE reduced the gap from 22.33 to 19.97 percentage points after adult-only fine-tuning and from 11.05 to 10.69-10.78 points after child-speech adaptation. It also lowered adult and child WER. Part of this advantage emerged before any child speech audio was used, suggesting an architectural contribution beyond data exposure. In the multilingual experiment, the benefit did not generalise uniformly. MoE reduced the gap significantly only for English. Dutch showed a smaller, non-significant reduction. German showed an increase in child WER and a significant widening of the adult-child gap, despite improved adult WER. The Polish results were difficult to interpret, as the corpus contained only isolated words and children with sigmatism. A routing analysis of the English models found age-correlated differences in early and late encoder regions after child-speech fine-tuning. The experts that adult and child speech were routed to produced distinct representations. An English domain-control analysis did not reproduce this pattern. Language was the dominant routing factor, and age-related specialisation was layered on top. Overall, a single shared MoE encoder can reduce the adult-child gap and develop age-related specialisation without demographic labels at inference time. Its effect depends on the coverage and domain match of the child speech data in each language. This makes the sparse MoE encoder a practical direction for more inclusive speech recognition: one model that reduces the adult-child gap without age labels, and whose routing can be examined to explain the improvement.

# Nomenclature

## Abbreviations

| Abbreviation | Definition                                                        |
|--------------|-------------------------------------------------------------------|
| AdamW        | Adam optimiser with decoupled weight decay                        |
| ASR          | Automatic speech recognition                                      |
| BPE          | Byte pair encoding                                                |
| CI           | Confidence interval                                               |
| CTC          | Connectionist temporal classification                             |
| CV           | Common Voice (crowdsourced multilingual speech corpus)            |
| DDP          | Distributed data parallel                                         |
| DNN          | Deep neural network                                               |
| E2E          | End-to-end                                                        |
| F0           | Fundamental frequency                                             |
| FFN          | Feed-forward network                                              |
| GELU         | Gaussian error linear unit                                        |
| GMM          | Gaussian mixture model                                            |
| HMM          | Hidden Markov model                                               |
| IPA          | International Phonetic Alphabet                                   |
| JASMIN       | Spoken Dutch Corpus extension for underrepresented speaker groups |
| JSD          | Jensen–Shannon divergence                                         |
| KidsTALC     | German child speech corpus (Rumberg et al., 2022)                 |
| KL           | Kullback–Leibler divergence                                       |
| LB           | Load balancing                                                    |
| LoRA         | Low-rank adaptation                                               |
| LR-MoE       | Language-routing mixture of experts                               |
| MoE          | Mixture of experts                                                |
| MoLE         | Mixture of language experts                                       |
| MyST         | My Science Tutor (English child speech corpus)                    |
| PAVSig       | Polish audio-visual sigmatism corpus                              |
| RNN-T        | Recurrent neural network transducer                               |
| SpecAugment  | Spectrogram augmentation                                          |
| WER          | Word error rate                                                   |
| YODAS2       | YouTube-Oriented Dataset for Audio and Speech, version 2          |

## Symbols

| Symbol                    | Definition                                                            |
|---------------------------|-----------------------------------------------------------------------|
| $B$                       | Number of bootstrap resamples                                         |
| $c_e^{(g)}$               | Number of times expert $e$ is selected across all frames of group $g$ |
| $d$                       | Encoder hidden dimension                                              |
| $D$                       | Number of deletion errors in WER                                      |
| $E_i(\mathbf{x})$         | Output of expert $i$ for input token $\mathbf{x}$                     |
| $f_i$                     | Fraction of batch tokens routed to expert $i$                         |
| $F^{(g)}$                 | Total number of input frames for group $g$                            |
| $g_i(\mathbf{x})$         | Router weight assigned to expert $i$ for input $\mathbf{x}$           |
| $G2norm_{\text{diff}}$    | Group-to-norm bias difference; child WER minus adult WER (pp)         |
| $I$                       | Number of insertion errors in WER                                     |
| $k$                       | Number of experts selected per token (top- $k$ routing)               |
| $N$                       | Number of experts per MoE layer (also: total reference words in WER)  |
| $\mathbf{o}_i$            | Output vector of expert $i$                                           |
| $p_e^{(g)}$               | Routing distribution probability for expert $e$ in group $g$          |
| $P_i$                     | Average router probability for expert $i$ over a batch                |
| $q$                       | False discovery rate threshold (Benjamini–Hochberg)                   |
| $S$                       | Number of substitution errors in WER                                  |
| $\tilde{g}_i(\mathbf{x})$ | Renormalised router weight for a selected expert                      |
| $\mathbf{x}$              | Input token representation                                            |
| $\beta_1, \beta_2$        | AdamW momentum coefficients                                           |
| $\Delta_e$                | Expert preference delta; child minus adult load for expert $e$        |
| $\Delta G$                | Difference in adult–child WER gap between two models (pp)             |
| $\Delta \text{WER}$       | WER difference between two models (pp)                                |
| $\ell_e^{(g)}$            | Expert load; average expert assignments per input frame for group $g$ |
| $\lambda$                 | Load-balancing loss coefficient                                       |
| $\mathcal{L}_{\text{lb}}$ | Load-balancing auxiliary loss                                         |
| $\sigma$                  | Standard deviation of Gaussian noise added to router logits           |

# Contents

|                                                                    |            |
|--------------------------------------------------------------------|------------|
| <b>Preface</b>                                                     | <b>i</b>   |
| <b>Abstract</b>                                                    | <b>ii</b>  |
| <b>Nomenclature</b>                                                | <b>iii</b> |
| <b>1 Introduction</b>                                              | <b>1</b>   |
| 1.1 Research Questions . . . . .                                   | 4          |
| 1.2 Thesis Outline . . . . .                                       | 4          |
| <b>2 Background</b>                                                | <b>5</b>   |
| 2.1 Automatic Speech Recognition . . . . .                         | 5          |
| 2.1.1 End-to-End ASR with CTC . . . . .                            | 6          |
| 2.2 End-to-End ASR Model Architectures . . . . .                   | 6          |
| 2.2.1 Conformer and FastConformer . . . . .                        | 6          |
| 2.2.2 Mixture-of-Experts . . . . .                                 | 7          |
| 2.3 Multilingual Speech Recognition . . . . .                      | 9          |
| <b>3 Methodology and Experimental Setup</b>                        | <b>10</b>  |
| 3.1 Overview . . . . .                                             | 10         |
| 3.2 Datasets . . . . .                                             | 10         |
| 3.2.1 Adult speech corpora . . . . .                               | 10         |
| 3.2.2 Child speech corpora . . . . .                               | 11         |
| 3.2.3 Corpus statistics: speaker distribution and gender . . . . . | 12         |
| 3.3 Data Preparation . . . . .                                     | 13         |
| 3.3.1 Shared preprocessing pipeline . . . . .                      | 13         |
| 3.3.2 Corpus-specific preprocessing . . . . .                      | 13         |
| 3.3.3 Train, development, and test splits . . . . .                | 14         |
| 3.4 Tokenisation . . . . .                                         | 14         |
| 3.4.1 Shared subword vocabulary . . . . .                          | 14         |
| 3.4.2 Vocabulary size and configuration . . . . .                  | 14         |
| 3.5 Model Architectures . . . . .                                  | 15         |
| 3.5.1 Dense FastConformer-CTC baseline . . . . .                   | 15         |
| 3.5.2 Sparse MoE FastConformer . . . . .                           | 16         |
| 3.5.3 Parameter counts . . . . .                                   | 17         |
| 3.6 Experimental Setup . . . . .                                   | 17         |
| 3.6.1 Compared model configurations . . . . .                      | 17         |
| 3.6.2 Two-stage training protocol . . . . .                        | 17         |
| 3.7 Training Setup . . . . .                                       | 18         |
| 3.7.1 Checkpoint initialisation . . . . .                          | 18         |
| 3.7.2 Optimisation settings . . . . .                              | 20         |
| 3.7.3 Stage 1: Adult speech fine-tuning . . . . .                  | 20         |
| 3.7.4 Stage 2: Child-speech fine-tuning . . . . .                  | 20         |
| 3.8 Evaluation Metrics . . . . .                                   | 21         |
| 3.8.1 Word error rate . . . . .                                    | 21         |
| 3.8.2 Performance disparity metric . . . . .                       | 21         |
| 3.8.3 Statistical significance testing . . . . .                   | 21         |
| 3.9 Experiments . . . . .                                          | 22         |
| 3.9.1 Experiment 1: English adult-child WER gap . . . . .          | 22         |
| 3.9.2 Experiment 2: Multilingual generalisation . . . . .          | 22         |
| 3.9.3 Experiment 3: Expert routing analysis . . . . .              | 22         |

|                                                                                     |           |
|-------------------------------------------------------------------------------------|-----------|
| 3.10 Routing Analysis . . . . .                                                     | 23        |
| 3.10.1 Analysis overview . . . . .                                                  | 23        |
| 3.10.2 Shared routing extraction and metrics . . . . .                              | 24        |
| 3.10.3 English models routing analysis . . . . .                                    | 26        |
| 3.10.4 Multilingual models routing analysis . . . . .                               | 27        |
| <b>4 Results</b>                                                                    | <b>29</b> |
| 4.1 Experiment 1: English Adult-Child WER Gap (RQ1) . . . . .                       | 29        |
| 4.2 Experiment 2: Multilingual Generalisation (RQ2) . . . . .                       | 30        |
| 4.3 Experiment 3: Expert Routing Analysis (RQ3) . . . . .                           | 32        |
| 4.3.1 English Routing Analysis . . . . .                                            | 32        |
| 4.3.2 Multilingual Routing Analysis . . . . .                                       | 39        |
| <b>5 Discussion</b>                                                                 | <b>47</b> |
| 5.1 Discussion of the results . . . . .                                             | 47        |
| 5.1.1 Does the MoE encoder reduce the adult-child gap in English ASR? (RQ1) . . . . | 47        |
| 5.1.2 Does the MoE encoder reduce the adult-child gap across languages? (RQ2) . .   | 48        |
| 5.1.3 Do experts specialise without demographic supervision? (RQ3) . . . . .        | 49        |
| 5.2 Limitations . . . . .                                                           | 51        |
| 5.2.1 Data limitations . . . . .                                                    | 51        |
| 5.2.2 Architectural and training-run limitations . . . . .                          | 52        |
| 5.2.3 Pathological child speech and matched controls . . . . .                      | 52        |
| 5.2.4 Performance disparity metric . . . . .                                        | 52        |
| 5.3 Social implications . . . . .                                                   | 53        |
| 5.4 Future Work . . . . .                                                           | 54        |
| 5.5 Contributions . . . . .                                                         | 55        |
| <b>6 Conclusion</b>                                                                 | <b>56</b> |
| <b>References</b>                                                                   | <b>58</b> |
| <b>A Acknowledgements</b>                                                           | <b>63</b> |
| A.1 Acknowledgements . . . . .                                                      | 63        |
| A.2 Acknowledgement of AI Tool Usage . . . . .                                      | 63        |

# 1

## Introduction

Automatic speech recognition (ASR) is a technology that transcribes spoken audio into written text. Over the years, ASR systems have improved substantially, driven by end-to-end neural architectures that extract acoustic and linguistic representations jointly from large amounts of transcribed speech [1], [2]. State-of-the-art systems achieve word error rates (WER) below 2% on standard English benchmarks [2], [3], and speech technology is now widely deployed in everyday tools such as voice assistants and automated transcription tools.

Despite these advances, improvements in ASR systems are not distributed uniformly. Systems consistently have higher error rates for atypical speakers, including children, elderly individuals, non-native speakers, and individuals with speech impairments, compared to typical adult native speakers [4], [5], [6], [7], [8]. This difference is largely linked to the dominance of adult, native, standard speech in large training corpora, which biases the acoustic and linguistic patterns of ASR models toward adult native speech and makes them less robust to the acoustic and pronunciation variability found in other speaker groups [5], [9], [10]. As a result, performance gaps between speaker groups persist even as overall accuracy increases, and optimising for aggregate benchmark scores can widen them further, since improvements concentrate on patterns already well-covered in training data [4], [7], [11]. When ASR serves as an interface in education, healthcare, or public services, these gaps directly affect accessibility [11].

Transcribing child speech remains a challenge for ASR: word error rates for children are often 30-50% on systems that achieve high performance on adult speech [6], [12]. These differences have acoustic foundations. Children have shorter vocal tracts than adults, which raises formant frequencies and shifts the spectral distribution of vowels far from the range that adult-trained models are optimised for [13]. Children produce higher and more variable fundamental frequency (F0) and phonological patterns that differ from adult patterns, such as child-specific simplification processes in fricatives, stops, and consonant clusters, and articulate less precisely [6], [14], [15]. Their speech also shows greater acoustic and linguistic variability than adult speech, reflecting ongoing development in articulation, speech motor control, and language use [6], [10], [16]. Unlike in adult training corpora, children produce more disfluencies, such as repetitions, fillers, and false starts [10]. Speaking rate and other interactional parameters vary with age, which has implications for acoustic and linguistic modelling in ASR systems for children [14]. Together, these developmental characteristics make child speech harder for ASR to transcribe accurately [6], [10], [14]. Children are frequent users of speech technology in education and everyday devices. Yet age-related performance disparities in ASR remain comparatively underexplored, particularly in multilingual settings [4].

Standard approaches to reducing the adult-child performance gap include adapting adult-trained models to child speech, typically through transfer learning or fine-tuning on child speech data [5], [10]. These methods increase child recognition performance but have two main limitations. First, they require child speech data, which is limited in quantity, especially for languages other than English [5], [10], [17]. Second, adaptation methods can introduce trade-offs between speaker groups: improving

recognition for one group can hurt performance on another. For example, fine-tuning to reduce errors on non-native accented speech came at the cost of native-speaker performance [7]. A similar trade-off may therefore also arise when adapting adult-trained ASR models toward child speech. More generally, this approach does not scale once variability sources combine. A non-native-accented child, for example, would need a model fine-tuned for that specific combination, and the number of combinations grows quickly as more speaker attributes are considered [10]. A model that handles both adult and child speech well within a single shared architecture, without requiring explicit speaker-type labels at inference time, would avoid the scaling and trade-off problems of per-group fine-tuning.

The adult-child performance gap is not limited to English; it becomes more prominent in multilingual settings. Many of the differences between child and adult speech come from general developmental factors, such as vocal-tract growth and developing speech motor control, so they are expected to affect ASR across languages [13], [16]. At the same time, these differences can result in distinct variability per language, depending on each language’s phonological structure. English, Dutch, and German are closely related Germanic languages, while Polish belongs to the Slavic family and has a different phonological structure, including a rich set of sibilant contrasts [15], [18]. These four languages were selected because they span two language families and because annotated child speech corpora exist for all of them. While multilingual ASR models are designed to handle variability across languages, they do not account for demographic variability within languages, such as the adult-child gap. In practice, multilingual training corpora are also more skewed toward standard adult speech than monolingual ones, which compounds the challenge [19]. This four-language setting within a single model, therefore, makes it possible to assess whether the MoE architecture can address both cross-language and age-related variability at once.

The Mixture-of-Experts (MoE) architecture, originally proposed as a way to divide a complex learning task among specialist sub-networks [20], offers a potential solution to this problem. In the MoE architecture, multiple expert networks each handle a particular portion of the input distribution. A router selects the most relevant expert(s) for each input, contrasting with a standard dense network, where every parameter is used for every input. The sparse MoE variant narrows this further, routing each input to only part of the experts; the remaining experts stay in memory but are not activated. This routing can capture latent structure in the data, such as differences between speakers or acoustic patterns, without requiring explicit labels, while the total computation per input stays low [21]. In ASR, sparse MoE layers within a Conformer encoder have been shown to improve multilingual recognition [19], [22], [23]. Routing choices assign different experts to different languages, without requiring any explicit language label at inference time. The resulting specialisation reduces word error rates relative to a dense model that has an equal number of active parameters [19], [22], [23].

MoE has a long history in speech processing, dating back to the 1990s. Peng et al. [24] applied a hierarchical MoE model to vowel recognition with multiple speakers, demonstrating that the router could discover speaker-specific routing patterns without any explicit speaker identity supervision. This early result showed the suitability of MoE-style modular models for speech recognition tasks involving speaker variability. More recently, SpeechMoE [25] scaled MoE to large acoustic models by replacing feed-forward network (FFN) layers in a Transformer encoder with dynamically-routed expert layers. The model achieved consistent improvements over dense baselines on Mandarin ASR benchmarks, with routing patterns showing data-driven phonetic and speaker-level specialisation [25].

Several recent studies have applied MoE to multilingual ASR. Gaur et al. [26] introduced Mixture of Informed Experts, in which per-language expert parameters are protected from gradients produced by other languages, reducing cross-language interference in an RNN-Transducer-based multilingual model. Kwon and Chung [23] proposed the Mixture of Language Experts (MoLE), where an utterance-level language gating network activates a language-specific expert and combines it with a language-agnostic expert according to the confidence of the gating decision. The gating network is trained with a language representation loss, which encourages utterances from the same language to be assigned consistently to the same expert. Wang et al. [19] proposed the Language-Routing Mixture of Experts (LR-MoE), which jointly trains a frame-level language identification network as the shared router across MoE layers, guiding language-specific expert activation for multilingual and code-switching ASR. Taken together, these studies show that MoE-based multilingual ASR has mainly used expert specialisation to model variation that is explicitly or structurally tied to language identity: through language-specific

gradient blocking, language-representation-guided expert selection, or a language-identification-based router. This is relevant for this thesis because it raises a related but different question: whether sparse MoE routing can also show systematic specialisation for speaker-group differences, such as adult versus child speech, without being given age labels as explicit routing supervision.

Hu et al. [22] implemented this design, replacing the FFN layers in a Conformer encoder with sparse MoE layers using top-2 routing, which inspired the architecture used in this thesis. Evaluated on 12 languages, it achieved an average 11.9% relative WER improvement over the dense multilingual baseline without requiring language identity at inference time, matching the performance of adapter-based models that do require language labels. This result showed that implicit routing in a Conformer-based multilingual model can be effective, but it did not consider child speech or age-related performance disparities.

MoE has also been used for other forms of speaker-level variability, beyond language. Perez et al. [27] used a mixture of intelligibility experts for aphasic speech recognition, where different experts specialised in different severities of speech impairment. Zhao et al. [28] proposed SAML, a mixture of LoRA experts for speaker-adaptive ASR that achieved substantial WER reductions with a sevenfold reduction in model size. These results suggest that MoE specialisation is not limited to language identity and can adapt to other speaker characteristics, which supports the idea that the same architecture could specialise for age-related speech differences.

Motivated by prior MoE-Conformer and shared-routing ASR models, this thesis studies a different setting: age-related variability in adult and child speech. The model used here is not a direct reproduction of Hu et al.'s MoE-Conformer system [22] or Gu et al.'s Transformer-based Omni-Router system [29]. Instead, this thesis implements a sparse MoE variant of FastConformer-CTC in NVIDIA NeMo. FastConformer is a more compute-efficient variant of the Conformer encoder that uses more aggressive input downsampling and is paired here with a CTC decoding head [3]; the full architectural specification is given in Section 3.5. In each Conformer block, the first feed-forward sublayer is removed and the second is replaced by a top-2 sparse MoE feed-forward module routed by a shared OmniRouter across encoder layers. Unlike language-routed MoE systems [19], [23], this router is not given age, language, or domain labels.

This design is motivated by the acoustic differences between adult and child speech described above. Because these differences are encoded in the same encoder representations used by the router, sparse expert routing may allow adult and child speech to activate different experts without being given age labels explicitly. The contribution is therefore not the invention of MoE routing itself, but its adaptation to a FastConformer-CTC adult-child ASR setting and the empirical analysis of whether age-related expert-routing patterns emerge without explicit age supervision.

This thesis investigates whether integrating sparse MoE layers into a FastConformer-CTC encoder affects the adult-child WER gap, both monolingually in English and multilingually across English, Dutch, German, and Polish. It further examines whether adult and child speech elicit different expert-routing patterns without explicit age labels, and whether the experts that receive this age-differentiated routing compute functionally distinct representations. This thesis addresses these questions by comparing MoE models with dense baselines and by analysing routing behaviour on held-out adult and child speech.

## 1.1. Research Questions

This thesis addresses three open questions.

- RQ1** Does integrating sparse Mixture-of-Experts layers into a FastConformer-CTC encoder reduce the Word Error Rate (WER) gap between adult and child speakers compared to a dense FastConformer baseline, while not degrading adult-speaker performance?
- RQ2** When trained on multilingual adult and child speech in English, Dutch, German, and Polish, how does the MoE encoder affect the adult-child WER gap relative to a dense baseline, and does this effect vary across languages?
- RQ3** Without explicit age labels, do routing distributions in a trained MoE encoder differ between adult and child speech within the same language, and do the experts that receive age-differentiated routing compute functionally distinct representations?

## 1.2. Thesis Outline

Chapter 2 provides the background and related work, covering automatic speech recognition and the CTC training objective, the FastConformer and Mixture-of-Experts encoder architectures used in this thesis, and multilingual ASR strategies such as shared-encoder training and data imbalance. Chapter 3 describes the proposed method and experimental design: the FastConformer-CTC baseline, the sparse MoE integration, the multilingual BPE tokenizer, the datasets, the training pipeline, and the evaluation protocol; its Experiments section is organised around the three research questions, with dedicated subsections for English experiments (RQ1), multilingual experiments (RQ2), and routing analysis (RQ3). Chapter 4 presents the results in the same order: English adult-child gap (RQ1), multilingual generalisation (RQ2), and expert routing behaviour (RQ3). Chapter 5 summarises and interprets the findings for each research question, discusses limitations, describes future work, and states the thesis's contributions. Chapter 6 concludes the thesis.

# 2

## Background

This chapter reviews the technical foundations and prior work relevant to this thesis. Section 2.1 introduces automatic speech recognition, covering its history and the CTC training objective. Section 2.2 describes the encoder architectures used in this thesis: the FastConformer baseline and the Mixture-of-Experts variant. Section 2.3 covers multilingual ASR strategies, including shared-encoder training and data imbalance.

### 2.1. Automatic Speech Recognition

Automatic Speech Recognition (ASR) is the task of mapping a continuous acoustic signal to a sequence of discrete output tokens, such as words or subwords [30], [31]. The task is inherently difficult because the same spoken word produces different acoustic signals depending on a speaker’s age, gender, accent, and speaking rate [32]. In addition, the output space is open-ended: since the set of feasible sentences in a language is infinite, ASR is formulated as a search over a discrete label space rather than a standard classification problem.

Before any modelling can take place, the raw waveform must be converted into a compact numerical representation. The signal is split into short, overlapping frames, and for each frame, the frequency spectrum is passed through a bank of triangular filters spaced according to the mel scale, a perceptual scale that approximates how humans discriminate pitch: it allocates more resolution to lower frequencies, where speech carries most of its phonetic content, and less to higher frequencies [33]. Taking the logarithm of each filter’s energy then compresses the dynamic range, giving the log-mel filterbank features used as input to the encoder architectures described in this thesis [33]. This turns thousands of raw audio samples per second into a much shorter sequence of feature vectors, one per frame, while retaining the acoustic detail needed for phonetic discrimination.

Early ASR systems used Hidden Markov Models to relate discrete phonetic units with continuous acoustic observations [31], [34]. Acoustic likelihoods were typically modelled using Gaussian Mixture Models, which captured the multimodal distribution of speech sounds across speakers and recording conditions [35]. However, GMMs represented acoustic variation through relatively simple Gaussian components and therefore had limited capacity to model the complex, nonlinear relationships between acoustic features, phonetic context, speaker characteristics, and noise conditions [36]. As deep learning matured, researchers replaced the GMM components with deep neural networks, which could learn more expressive hierarchical representations of speech and provide stronger discriminative estimates of acoustic states [36]. Hybrid systems still trained each component separately: the acoustic model learned to predict HMM states from frame-level alignments produced by a previously trained GMM-HMM system, while the pronunciation lexicon and language model were estimated independently and combined with the acoustic model only at decoding time [31], [36]. Because these components were not optimised jointly, errors in the alignments or in any single component could not be corrected by the others and propagated through to the final output [37], [38]. This limitation motivated end-to-end (E2E) approaches that jointly optimised all components within a single differentiable model [37], [39].

Modern E2E ASR encoders use deep neural networks, trained with gradient-based optimisation. Deep encoders have been shown to capture phonetic and fine-grained acoustic detail in lower layers, while higher layers encode more abstract, task-specific representations [40]. Most modern end-to-end systems, such as Conformer-based encoders [2], represent text as sequences of subword units rather than full words. This avoids a large, fixed vocabulary and lets systems represent unseen words as combinations of known pieces. Byte Pair Encoding (BPE) [41] is a widely used method for obtaining these units. It begins with individual characters and iteratively merges the most frequent adjacent symbol pair until a target vocabulary size is reached. Common morphemes become single tokens, while rare words are segmented into smaller pieces.

### 2.1.1. End-to-End ASR with CTC

A major challenge in E2E ASR is that the input sequence is much longer than the output sequence. The alignment between individual frames and output tokens is unknown at training time. Graves et al. [1] introduced Connectionist Temporal Classification (CTC), which solves this by summing over all valid input-output alignments during training, rather than committing to a single one. CTC adds a special blank symbol to the output vocabulary that represents "no output" at a given frame. The training objective maximises the total probability of the correct label sequence summed across all possible alignments that collapse to it. This procedure allows the encoder to be trained directly from paired audio and text transcripts, without frame-level labels indicating which token each frame corresponds to, and without the separate forced-alignment step that hybrid GMM-HMM systems require to produce those labels.

At inference time, the model emits a sequence of tokens and blanks at each input frame step. The final predicted transcript is created by collapsing consecutive repeated tokens and removing all blanks. For example, the raw frame-level output `-c-a-tt-` becomes "cat". Since CTC predictions are conditioned on the input representation rather than on previously emitted output tokens, CTC decoding is non-autoregressive under its conditional-independence assumption [1].

## 2.2. End-to-End ASR Model Architectures

This section describes the encoder architectures used in this thesis: the FastConformer, which forms the dense baseline, and the Mixture-of-Experts variant built on top of it.

### 2.2.1. Conformer and FastConformer

The baseline encoder used in this thesis is the FastConformer, which builds on the Transformer and Conformer architectures. Each step in this progression adds one specific capability to its predecessor: self-attention in the Transformer, convolution for local acoustic detail in the Conformer, and more efficient input subsampling in the FastConformer. Understanding what each of these three additions contributes is necessary for the encoder design described in Chapter 3.

The Transformer architecture [42] uses multi-head self-attention to relate any two positions in a sequence directly, regardless of distance. Applied to speech, however, purely attention-based encoders can miss fine-grained local patterns such as phoneme boundaries and coarticulation cues, which are important for accurate transcription.

The Conformer [2] addresses this by combining self-attention with depthwise convolution within each encoder block. As shown in Figure 2.1, the block follows a Macaron-style arrangement in which the attention and convolution modules are surrounded by two half-step feed-forward layers, each contributing half of its output weight. The convolution module captures local acoustic patterns within a fixed kernel width, while the attention module attends to the full input context. This combination gives the encoder both local and global acoustic sensitivity, and the Conformer outperforms purely attention-based and convolutional models on the LibriSpeech benchmark [2].

FastConformer [3] improves on the Conformer by subsampling the input sequence by a factor of 8 using depthwise-strided convolutions before the encoder blocks, compared to  $4\times$  in the standard Conformer. The shorter sequence reduces self-attention costs; Rekish et al. reported a roughly 2.8-fold speedup with comparable accuracy in their encoder-throughput benchmark.

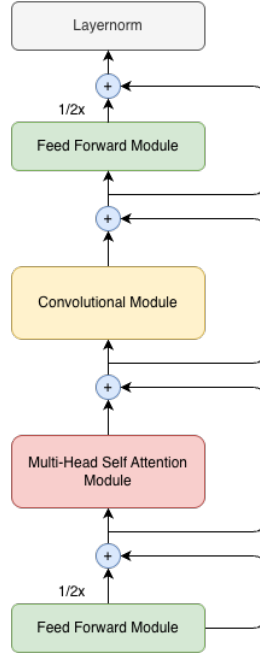


**Figure 2.1:** Conformer block used in the FastConformer encoder, based on [2], [3]. Each block uses a Macaron-style arrangement: a half-step feed-forward layer, multi-head self-attention, a depthwise convolution module, and a second half-step feed-forward layer, all with residual connections. The full encoder stacks  $N$  such blocks after an  $8\times$  depthwise-strided subsampling stage.

### 2.2.2. Mixture-of-Experts

In this thesis, Mixture-of-Experts (MoE) is applied by replacing the feed-forward sublayers in the FastConformer encoder with MoE layers. Each MoE layer contains  $N$  independent feed-forward expert networks and a router that selects which  $k$  of those experts process each input token. The introduction motivated why this mechanism is a natural candidate for handling adult-child and cross-language variability; this section explains how it works.

#### Sparse MoE architectures

Jacobs et al. [20] originally proposed the MoE framework as a modular approach to supervised learning. The core idea is to divide a complex prediction problem among a set of expert networks, each specialising in a subset of the input space, guided by a router that determines how much each expert contributes to a given input. All expert outputs are combined as a weighted sum, with the weights determined by the router. Training uses a competitive learning signal that encourages each expert to specialise in inputs it handles well, thereby partitioning the problem space in a data-driven way. Jordan and Jacobs [43] extended this to a hierarchical structure in which experts are themselves composed of sub-experts, and derived an Expectation-Maximisation algorithm that interprets the router weights as posterior probabilities over expert assignments.

In the original version of the Mixture-of-Experts setup, every expert processes every input. This means that using more experts also makes the model more expensive to run. Sparse MoE avoids this problem by using only a few experts per input. Usually, the model activates the top- $k$  experts, where  $k$  is often 1 or 2. The router still gives a score to every expert, but only the highest-scoring experts receive the input; the rest are skipped. This is the main trade-off: adding more experts increases the model's total capacity and allows greater specialisation, but each input is only processed by  $k$  of them. Therefore, the expert-network computation per token does not grow proportionally with the total number of expert parameters. Masoudnia and Ebrahimpour [21] give a wider overview of MoE variants and how they assign inputs to experts. Their overview distinguishes between methods in which experts learn to specialise through competition and methods in which the input space is divided in advance, for example,

through clustering.

In Transformer-based architectures, sparse MoE is commonly implemented by replacing the position-wise feed-forward network (FFN) sublayer in selected encoder or decoder blocks with an MoE-FFN [22], [44]. Each expert in the MoE-FFN is an independent FFN module with its own parameters, while a router selects the top- $k$  experts used to process each token. This design is important because it decouples total model capacity from per-token computation: increasing the number of experts increases the total number of parameters, but each token is processed by only  $k$  experts.

#### Routing mechanisms

Routing assigns each input token to a subset of the available experts. The simplest formulation is dense weighted routing: a linear projection maps the token representation  $\mathbf{x} \in \mathbb{R}^d$  to a score vector, and a softmax over all  $N$  experts gives routing weights  $g_i(\mathbf{x})$  that sum to one. The layer output is then just a weighted sum of all expert outputs:

$$\text{MoE}(\mathbf{x}) = \sum_{i=1}^N g_i(\mathbf{x}) E_i(\mathbf{x}) \quad (2.1)$$

Every expert is evaluated for every token, so this is equivalent to a weighted ensemble, and the computational cost rises linearly with  $N$ .

Sparse top- $k$  routing removes this cost by selecting only the  $k$  highest-scoring experts and setting all other weights to zero. The selected weights are renormalised to sum to one before the weighted sum is computed:

$$\text{MoE}(\mathbf{x}) = \sum_{i \in \text{TopK}(k)} \tilde{g}_i(\mathbf{x}) E_i(\mathbf{x}) \quad (2.2)$$

The total computation per token is now fixed at  $k$  expert evaluations regardless of  $N$ , so adding more experts increases model capacity without increasing inference cost. This thesis uses top- $k = 2$  routing for all MoE models.

One practical problem with sparse Mixture-of-Experts models is expert imbalance or expert collapse, where the router sends most inputs to a small number of experts while other experts receive little training signal and remain underused [44], [45]. A standard fix for this problem in the prior literature is to add an auxiliary load-balancing loss to the main training objective. This extra loss encourages the router to spread tokens more evenly across all experts [44], [45].

For a batch of tokens, the load balancing loss is:

$$\mathcal{L}_{\text{lb}} = \lambda \cdot N \sum_{i=1}^N f_i P_i \quad (2.3)$$

where  $N$  is the number of experts and  $\lambda$  controls how strongly the model is encouraged to balance the load. For each expert  $i$ , the term  $f_i$  measures how many tokens were actually sent to that expert, while  $P_i$  measures how much probability the router assigned to that expert on average.

In simple terms, this loss penalises the model when some experts are used much more than others. The loss is lowest when all experts are used equally, meaning each expert receives about  $1/N$  of the tokens.

The term  $f_i$  records the actual routing choices, while  $P_i$  allows gradients to flow back through the router because it is based on the router's soft probabilities. This allows training to adjust the router's behaviour by gradient updates. The formula above follows [44], [45]; the exact definition of  $f_i$  and how it scales with top- $k$  routing depends on the implementation. The formulation used in this thesis is given in Chapter 3.

Another technique for avoiding expert collapse is to add Gaussian noise to the router logits before selecting the top- $k$  experts. This noise makes the router explore different experts more often at the beginning of training, instead of immediately committing to only a few experts.

Regardless of the specific routing design, the router’s only input is the encoder’s output vector for each token, a fixed-size numerical representation of the acoustic signal at that point in the sequence. It has no access to speaker age, language identity, or any other metadata. Any pattern in routing that correlates with such properties cannot arise from direct access to these labels. However, the encoder representation may also encode speaker characteristics, phonetic and lexical content, speaking style, and recording conditions. Routing patterns should therefore be interpreted as associations with input properties, not as proof that the router has isolated a specific acoustic characteristic.

MoE models differ in whether each layer maintains its own router parameters or whether a single set of router parameters is shared across all MoE layers. In standard per-layer routing, each layer applies an independent router to its own hidden representations, and expert assignments can vary substantially from layer to layer. The Omni-Router [29] instead shares the router weight matrix across all MoE layers. Because each layer still applies this shared router to its own hidden representation, a token may be routed to different experts at different depths. Sharing router parameters encourages more coherent routing decisions across layers without constraining the assignments to be identical, which Gu et al. argue promotes inter-layer coordination and expert specialisation [29].

## 2.3. Multilingual Speech Recognition

Multilingual ASR systems process speech from multiple languages within a single model. The main challenges are managing cross-language interference in the shared encoder and handling imbalanced training data across languages [26], [46]. Following common usage in the ASR literature, a language’s resource level refers to how much transcribed speech data is available for training, not to the number of people who speak it: a *high-resource* language such as English has large transcribed corpora, while a *low-resource* language has comparatively little [26], [46].

The simplest approach to multilingual ASR is a fully shared model trained on data from all languages simultaneously, using a common encoder and a shared output vocabulary [46]. The vocabulary is typically a multilingual BPE tokenizer [41], [46] trained on text from all languages, which clusters frequently co-occurring character sequences into subword units. For languages with overlapping character sets, such as English, Dutch, and German, shared subwords can capture surface-form similarities across languages [41], [46]. This architecture benefits low-resource languages by exposing their data to representations learned from higher-resource languages, but it can also introduce interference: gradients from different languages may conflict in the shared encoder, lowering performance on individual languages, particularly high-resource ones [26].

One way to reduce cross-language interference is to give each language its own model components while still sharing most of the model [46], [47]. For example, some approaches use separate output heads for different languages [46]. Others add a language ID token at the start of the input, or use small language-specific adapter layers [47], [48]. In both cases, the language label must be known at inference time [48]. If the language is not known, an automatic language identification system must be used first [48]. This adds an extra step, increases latency, and can introduce mistakes if the language is identified incorrectly [48]. Mixture-of-Experts architectures, introduced in Section 2.2.2, provide another option. Instead of requiring an explicit language label, the router decides which parts of the model to use for each input. This enables the model to adapt to different languages automatically at inference time [19], [22], [23].

A further practical challenge in multilingual training is data imbalance. Languages typically contribute different amounts of training data [26], [47]. If all data is simply combined and sampled in proportion to its size, the model sees far more examples from high-resource languages per training epoch than from low-resource ones [47], [49]. Equal language probability sampling, which draws a fixed fraction of each training batch from each language regardless of its total dataset size, is one approach to this imbalance [49], [50]. This means that low-resource languages are repeated across training epochs, while high-resource languages are downsampled, giving each language more comparable influence on the learned representations [49], [50].

# Methodology and Experimental Setup

## 3.1. Overview

This chapter describes the methodology and experimental design used to address the three research questions stated in Chapter 1. Two encoder architectures are compared: a dense FastConformer-CTC baseline and a sparse Mixture-of-Experts (MoE) variant of the same encoder. Both are initialised from the same pre-trained NVIDIA FastConformer checkpoint and fine-tuned in two stages: first on adult speech, then on a mixture of adult and child speech.

Three experiments are conducted. Experiment 1 tests whether MoE reduces the adult-child word error rate gap in English without degrading adult performance. Experiment 2 tests how that effect changes across Dutch, German, and Polish. Experiment 3 analyses the routing behaviour of the trained MoE encoder to determine whether experts specialise in adult or child speech without explicit age supervision.

The datasets used are introduced in Section 3.2, followed by data preparation in Section 3.3 and tokenisation in Section 3.4. The model architectures are defined in Section 3.5. Detailed experiment configurations, including the full model comparison table and training protocol, are given in Section 3.6. Training hyperparameters are in Section 3.7, evaluation metrics in Section 3.8, and the routing analysis methodology in Section 3.10.

## 3.2. Datasets

This section describes the datasets used in the experiments. This work focuses on age inclusivity, so datasets are divided into two groups: adult speech corpora, which are larger and more widely available, and child speech corpora, which are typically smaller and harder to access.

### 3.2.1. Adult speech corpora

#### LibriSpeech

LibriSpeech [51] is a corpus of approximately 960 hours of English read speech derived from public-domain audiobooks. The predefined training splits `train-clean-100` (100 h), `train-clean-360` (360 h), and `train-other-500` (500 h) are combined and used as the English training data in both experiments. During training, `dev-clean` and `dev-other` are combined to form the validation set. For evaluation, `test-clean` and `test-other` are combined to form the test set in the experiments.

#### Common Voice

Common Voice [52] is a crowdsourced multilingual speech corpus in which contributors record prompted sentences via a web interface. Each recording is validated by other contributors; only clips with more upvotes than downvotes are included in the official splits. Version 25.0 was used for Dutch, German, and Polish. All three subsets were preprocessed to 16 kHz mono WAV. Gender metadata is optionally self-reported: across the full release, gender information is missing for 66.5% of Dutch speakers, 77.2% of German speakers, and 79.2% of Polish speakers (see Table 3.3).

Together, LibriSpeech and Common Voice form the multilingual adult training corpus, which comprises four languages: English (960 h from the full LibriSpeech training set: `train-clean-100`, `train-clean-360`, and `train-other-500`), German (1,206 h from Common Voice 25.0 [52]), Dutch (88 h from Common Voice 25.0), and Polish (125 h from Common Voice 25.0). The three Common Voice languages were split speaker-independently, assigning 10% of speakers to the test partition, 5% to validation, and the remaining 85% to training; English instead uses the predefined LibriSpeech training, validation, and test splits described above. Because per-speaker contribution volume in Common Voice varies widely, a split with fewer speakers can still contain more audio; this is why the Dutch validation partition has more hours than the Dutch test partition in Table 3.1. Table 3.1 summarises the split statistics for each language.

**Table 3.1:** Multilingual adult training and evaluation dataset statistics. The Common Voice subsets are speaker-independently split into train, validation, and test partitions. LibriSpeech uses the predefined training, validation, and test splits, with LibriSpeech (LS) hours combined across `train-clean-100`, `train-clean-360`, and `train-other-500`. Gender figures are per-split speaker percentages.

| Language       | Source      | Split | Clips     | Hours   | Speakers | Gender (M% / F% / unsp.%) |
|----------------|-------------|-------|-----------|---------|----------|---------------------------|
| Dutch (NL)     | CV 25.0     | Train | 71,836    | 88.3    | 1,539    | 26.1 / 7.4 / 66.5         |
|                |             | Val   | 18,693    | 21.1    | 90       | 30.0 / 5.6 / 64.4         |
|                |             | Test  | 13,422    | 15.3    | 180      | 22.8 / 7.2 / 70.0         |
| German (DE)    | CV 25.0     | Train | 832,553   | 1,205.9 | 16,715   | 19.4 / 3.4 / 77.2         |
|                |             | Val   | 40,468    | 58.4    | 983      | 19.2 / 2.8 / 77.9         |
|                |             | Test  | 76,677    | 107.8   | 1,966    | 18.9 / 3.1 / 78.0         |
| Polish (PL)    | CV 25.0     | Train | 99,637    | 124.6   | 2,834    | 18.7 / 2.2 / 79.2         |
|                |             | Val   | 8,620     | 11.3    | 166      | 23.5 / 2.4 / 74.1         |
|                |             | Test  | 30,035    | 37.1    | 333      | 23.4 / 3.3 / 73.3         |
| English (EN)   | LibriSpeech | Train | 281,241   | 961.1   | 2,338    | 51.9 / 48.1 / 0.0         |
|                |             | Val   | 5,567     | 10.5    | 73       | 50.7 / 49.3 / 0.0         |
|                |             | Test  | 5,559     | 10.7    | 73       | 49.3 / 50.7 / 0.0         |
| Total Training |             |       | 1,285,267 | 2,379.9 | 23,426   |                           |

The four languages contribute unequal amounts of data: English 960 h, German 1,206 h, Polish 125 h, and Dutch 88 h. Sampling in proportion to these totals would cause the model to train predominantly on German and English, and only rarely on Dutch or Polish; the sampling strategy used to address this is described in Section 3.7.3.

#### YODAS2

YODAS2 (YouTube-Oriented Dataset for Audio and Speech, version 2) [53] is a large-scale corpus of English YouTube speech with automatic transcriptions filtered by transcription quality. A subset of the English portion at quality bucket 0 is used for evaluation only, in the routing analysis domain control experiment (Section 3.10.3), to test whether routing differences between LibriSpeech and MyST are driven by recording conditions rather than speaker age.

### 3.2.2. Child speech corpora

Four child speech corpora are used: MyST (English), JASMIN (Dutch), KidsTALC (German), and PAVSig (Polish). Their descriptions are given below; corpus-specific preprocessing steps are described in Section 3.3.2. Table 3.2 summarises the resulting dataset statistics.

#### MyST (English)

The My Science Tutor (MyST) corpus [54] is a large corpus of children’s conversational speech collected during spoken interactions between elementary-school students and a virtual science tutor. The corpus contains approximately 393 hours of speech from 1,371 students in grades 3-5, comprising 228,874 utterances across 10,496 tutoring sessions. The recordings were collected as part of a long-running educational project in which students interacted one-on-one with a virtual tutor about classroom science

topics. Each student's turn was recorded as a separate audio file, and no identifying information was stored with the data; students were represented only through anonymised school and student codes.

Speaker identity was determined from the student ID encoded in the directory and session structure. The corpus provides predefined `training`, `development`, and `test` partitions with no speaker overlap across splits, and this partitioning is used without modification. After the quality filtering and session merging described in Section 3.3.2, the subset used in this thesis is reduced to 566 speakers, 65,885 utterances, and 199.60 hours of speech (Table 3.2).

#### JASMIN (Dutch)

The JASMIN-CGN corpus [9] extends the Spoken Dutch Corpus with speech from underrepresented groups, including children, non-native speakers, and elderly speakers, recorded in read-speech and human-machine interaction settings. The child subset used in this thesis consists of 74 native Dutch speakers aged 6–13 (37 female, 37 male), providing 7.87 hours of speech across 13,762 utterances after preprocessing (Table 3.2). The corpus provides per-recording metadata including speaker group, age, gender, and dialect region.

For the experiments, only native Dutch child speech from the Netherlands was retained. Flemish recordings were excluded to keep the Dutch child data consistent with the adult Dutch evaluation data.

#### KidsTALC (German)

The KidsTALC corpus [17] is a German child speech corpus with speech from monolingual, typically developing children, elicited through picture-description, narrative, and conversational tasks in quiet home or school settings. The subset used in this thesis consists of 38 speakers aged 4–11 (19 female, 19 male), providing 8.84 hours of speech across 8,519 utterances after preprocessing (Table 3.2).

#### PAVSig (Polish)

The PAVSig corpus [18] is a Polish audio-visual child speech dataset designed to support computer-aided diagnosis of sigmatism, a speech sound disorder affecting sibilant production. Recordings were collected from 201 native Polish-speaking children aged 4;10–8;0 during scheduled speech therapy examinations; 61 are labelled as having typical articulation and 140 as having atypical articulation. Utterances are short isolated words and logotomes (pronounceable phoneme sequences with no standard orthographic form), not continuous speech; the full corpus contains 66,781 audio-video segments, which reduce to 65 unique word classes after orthographic mapping and preprocessing (Section 3.3.2). Because this is a small fixed-vocabulary corpus, the original corpus splits were replaced with a word- and speaker-disjoint re-split; details are given in Section 3.3.2.

The summary statistics of the child speech datasets are presented in Table 3.2.

**Table 3.2:** Child speech corpora after preprocessing.

| Corpus   | Lang. | Speakers | Utterances | Hours  | Age range |
|----------|-------|----------|------------|--------|-----------|
| MyST     | EN    | 566      | 65,885     | 199.60 | 8–11      |
| JASMIN   | NL    | 74       | 13,762     | 7.87   | 6–13      |
| KidsTALC | DE    | 38       | 8,519      | 8.84   | 4–11      |
| PAVSig   | PL    | 201      | 9,923      | 4.13   | 4–8       |

### 3.2.3. Corpus statistics: speaker distribution and gender

Gender distribution is relevant because reported gender is associated with systematic acoustic differences, including differences in fundamental frequency and vocal tract length. A strongly male-skewed corpus may therefore affect how well a model generalises across speaker groups. Table 3.3 shows that, among Common Voice speakers who self-reported a gender, male speakers greatly outnumber female speakers across all three subsets (19–26% labelled male vs 2–7% labelled female). However, the majority of speakers (67–79%) did not report a gender label, so this observed skew reflects only the subset who chose to disclose; the gender composition of the corpus as a whole cannot be reliably determined from this data. LibriSpeech, by contrast, has a near-balanced reported gender distribution (52% male, 48% female). The child corpora are also balanced by design, with a 50/50 split in JASMIN

and KidsTALC. Because the true gender composition of the Common Voice subsets is unknown, the possibility of an undetected gender skew is treated as a limitation when interpreting the results.

**Table 3.3:** Corpus-level statistics: unique speakers and gender distribution where available. Gender is self-reported for Common Voice; figures are per-speaker percentages across the full CV 25.0 release. Between 67% and 79% of Common Voice speakers carry no gender label (unspec.).

| Corpus                | Lang. | Speakers | Gender distribution                         |
|-----------------------|-------|----------|---------------------------------------------|
| LibriSpeech 360       | EN    | ~921     | M 52%, F 48%                                |
| LibriSpeech 100       | EN    | ~251     | M 52%, F 48%                                |
| LibriSpeech other-500 | EN    | ~1,166   | M 52%, F 48%                                |
| Common Voice DE       | DE    | 19,664   | M 19.4%, F 3.4%, other 0.03%, unspec. 77.2% |
| Common Voice PL       | PL    | 3,333    | M 18.7%, F 2.2%, other 0%, unspec. 79.2%    |
| Common Voice NL       | NL    | 1,809    | M 26.1%, F 7.4%, other 0%, unspec. 66.5%    |
| MyST                  | EN    | 566      | Not recorded                                |
| JASMIN                | NL    | 74       | M 50%, F 50%                                |
| KidsTALC              | DE    | 38       | M 50%, F 50%                                |
| PAVSig                | PL    | 201      | M 46.8%, F 53.2%                            |

### 3.3. Data Preparation

#### 3.3.1. Shared preprocessing pipeline

All corpora are processed through the following steps before training and evaluation.

1. **Audio normalisation.** All audio is resampled to 16 kHz mono.
2. **Duration filtering.** Utterances shorter than 0.5 s or longer than 40.0 s are excluded at training time.
3. **Empty transcription removal.** Utterances with empty text fields are discarded.
4. **Whitespace normalisation.** Any sequence of multiple spaces is reduced to one; leading and trailing spaces are stripped.

For training, 80-dimensional log-mel filterbank features are extracted from the normalised audio using a 25 ms analysis window and a 10 ms hop, as defined in the model configuration [3]. SpecAugment [55] data augmentation is applied during training with two frequency masks (maximum width 27 channels) and ten time masks (maximum width 5% of sequence length).

#### 3.3.2. Corpus-specific preprocessing

The steps below apply only where the corpus format requires them. All other corpora use the shared pipeline without modification.

**MyST.** Each transcription is normalised before use. Whisper’s English text normaliser standardises contractions, punctuation, and abbreviations; digit sequences are then converted to their spoken-word equivalents (e.g. 3 → *three*); and finally the text is lower-cased.

Three quality filters are applied to remove unreliable utterances. First, utterances whose transcription consists only of a non-speech label such as <silence> or <noise>, or that contain fewer than three words, are discarded. Second, recordings where more than 1% of audio samples reach near-maximum amplitude are discarded, as clipping at this level distorts the signal. Third, each utterance is independently transcribed using Whisper large; utterances where the word error rate between the Whisper hypothesis and the reference transcription exceeds 50% are discarded as likely mismatches between the audio and its label. Finally, short utterances from the same recording session are merged. Consecutive clips of up to 12 seconds are concatenated into a single clip, up to a maximum of 30 seconds. This reduces the number of very short segments while keeping utterances within the same session together.

**JASMIN.** Transcriptions contain corpus-specific non-speech markers and artefacts. Trailing punctuation is stripped; utterances containing truncation markers ( . . . ) are discarded, as incomplete words introduce label noise incompatible with CTC training; utterances consisting solely of non-speech tokens (ggg, xxx, yyy, uh, um) are removed; individual non-speech tokens within otherwise valid utterances are deleted. The existing train/development/test partitioning is preserved without modification.

**KidsTALC.** Original capitalisation is preserved during preprocessing, matching the casing convention used by the multilingual tokenizer (Section 3.4). The test split provided by the corpus carries no transcriptions, so the validation split is used as the test set instead.

**PAVSig.** The original transcriptions use IPA phonetic notation, which is incompatible with the orthographic BPE tokenizer. Because the corpus has a fixed, known vocabulary, IPA transcriptions are replaced with standard Polish orthographic forms via a lookup table, restoring diacritics where absent (e.g. *noz* → *nóž*). Logotomes are retained as written.

PAVSig contains only 65 unique word classes and 201 speakers, so using the original splits would allow the model to be evaluated on words and speakers seen during training, inflating test performance. Because this is a small fixed-vocabulary corpus, the corpus was re-split so that test words and test speakers are fully disjoint from training. Approximately 20% of words (13 word classes) and 20% of speakers (40 speakers) were designated as held-out test identities; a sample enters the test set only if it contains both a held-out word and a held-out speaker, preventing leakage through shared word or speaker identity. The resulting split contains 7,848 training samples, 1,569 validation samples, and 506 test samples. PAVSig recordings include both mono and multi-channel audio files. Multi-channel files are converted to mono by averaging across channels; mono files are unchanged.

### 3.3.3. Train, development, and test splits

Speaker-independent splits are used for all corpora. Where a corpus provides predefined splits, these are retained: LibriSpeech (Section 3.2.1), MyST (Section 3.2.2), and JASMIN (Section 3.2.2). The KidsTALC validation split is used as the test set (Section 3.2.2). Common Voice instead uses a custom 85%/5%/10% speaker-independent split (Section 3.2.1), and PAVSig uses a custom word- and speaker-disjoint split (Section 3.2.2).

## 3.4. Tokenisation

### 3.4.1. Shared subword vocabulary

All models use a SentencePiece Byte Pair Encoding (BPE) tokenizer [41], [56]. BPE is described in Section 2.1. Two distinct tokenizers are used in this thesis, one per experiment rather than one per language: a dedicated tokenizer for the English-only Experiment 1 models, and a single shared tokenizer for the four-language Experiment 2 models. This follows from the shared-encoder multilingual architecture introduced in Section 2.3: Experiment 2 trains one encoder with a single CTC output projection across English, Dutch, German, and Polish simultaneously, rather than separate output heads per language, so that projection requires one shared vocabulary covering all four languages.

The multilingual tokenizer is not reused for Experiment 1 either, because vocabulary size is sized to the number of languages actually present in an experiment (Section 3.4.2). Experiment 1 uses English data only, so its tokenizer is trained, and its vocabulary sized for English alone. Using the larger multilingual vocabulary instead would allocate a substantial share of subword units to German, Dutch, and Polish forms that never occur in Experiment 1’s data, without any benefit to the English-only task.

### 3.4.2. Vocabulary size and configuration

The two tokenizers apply different casing conventions. This reflects how systematically case functions as a grammatical marker in each language set.

**English tokenizer.** The English experiments use a BPE tokenizer with a vocabulary of 4,096 subword units trained on the training-set transcriptions of LibriSpeech (Section 3.2.1) and MyST (Section 3.2.2), with lower-casing applied. English capitalisation is largely restricted to sentence-initial position and

proper nouns rather than functioning as a systematic grammatical marker, and LibriSpeech’s official transcriptions are already provided in a single uppercase form with no case distinctions. Lower-casing therefore standardises the vocabulary rather than discarding a systematic orthographic signal. Validation and test transcriptions are not used during tokenizer training, so the vocabulary is constructed without any exposure to held-out text. Including MyST training transcriptions at tokenizer training time ensures that child-specific vocabulary and informal speech patterns fall within the vocabulary’s coverage.

**Multilingual tokenizer.** The multilingual experiments use a separate BPE tokenizer with a vocabulary of 16,384 subword units trained on the training-set transcriptions of LibriSpeech for English and Common Voice for German, Dutch, and Polish (Section 3.2.1), together with the training-set transcriptions from all four child speech corpora: MyST, JASMIN, KidsTALC, and PAVSig (Section 3.2.2). Validation and test transcriptions are excluded from tokenizer training. No lower-casing is applied, because German nouns are capitalised by convention: unlike English casing, this is a systematic, near-universal grammatical marker, so preserving it maintains a consistent orthographic distinction rather than fragmenting the vocabulary. The larger vocabulary accommodates the wider character and morpheme inventory across four languages. All four languages use the Latin alphabet. English, Dutch, and German share substantial character-level overlap, so BPE merges distribute across their shared subword patterns. Polish introduces diacritic characters (e.g., *ą, ę, ó, ś, ź*) absent from the other three languages. The larger vocabulary also partially compensates for the higher morphological complexity of Polish, allowing more inflectional forms to be represented as single tokens rather than multi-piece sequences. The chosen size corresponds to roughly 4,000 units per language, a design choice intended to limit subword fragmentation in the morphologically richer languages while keeping the shared CTC output projection at a manageable size; subword segmentation itself follows the BPE approach of [41].

Because both tokenizers include child-corpus training transcriptions, the Stage 1 adult-only models have text-level exposure to child speech through the subword vocabulary. This exposure is textual only: no child speech audio is used before Stage 2, so Stage 1 comparisons isolate acoustic generalisation to child speech.

## 3.5. Model Architectures

This section defines the two encoder architectures compared in this thesis: the dense FastConformer-CTC baseline and the sparse MoE-augmented variant. The study overview in Section 3.1 introduced both at a high level; this section gives the full technical specification. The experiment-specific model configurations, which models are trained and compared in each experiment, are listed in Section 3.6, after the architectures have been defined.

### 3.5.1. Dense FastConformer-CTC baseline

#### Architecture overview

The baseline is the Large FastConformer-CTC [3] configuration: 17 Conformer blocks with model dimension  $d = 512$ , 8 attention heads, and  $8\times$  depthwise-strided subsampling, giving approximately 110 million parameters as instantiated in this thesis.

The acoustic front-end extracts 80-dimensional log-mel filterbanks from 16 kHz audio with a 25 ms analysis window and a 10 ms frame shift. Before the Conformer blocks, a stack of depthwise separable convolutions with 256 channels reduces the feature sequence by a factor of 8, from roughly 100 frames per second to 12.5 frames per second.

#### Encoder design

Each Conformer block follows the Macaron-style arrangement described in Section 2.2.1, with residual connections and layer normalisation after each sublayer. Layer normalisation rescales each token’s activations to zero mean and unit variance across the feature dimension. The depthwise convolution sublayer uses a kernel size of 9.

Multi-head self-attention uses relative positional encoding with full context: each frame attends to the entire encoder sequence after subsampling, with no left or right restriction (context window  $[-1, -1]$ ).

### CTC decoder

The decoder is a single linear projection from  $d = 512$  to the vocabulary size, followed by log-softmax. The model is trained with the CTC objective introduced by [1] and described in Section 2.1.1. CTC is used rather than a transducer decoder [57] because it uses a simpler non-autoregressive decoder and avoids the additional prediction and joint networks required by transducer models, reducing the computational overhead during fine-tuning. This is useful given the modest size of the available child speech corpora. Because the decoder adds no network structure beyond the output projection, the architecture can be extended with MoE layers in the encoder without changing the decoder.

### 3.5.2. Sparse MoE FastConformer

The MoE model shares the FastConformer encoder structure with the dense baseline. The only structural difference is that the first feed-forward sublayer in each Conformer block is removed, and the second is replaced by a sparse MoE feed-forward layer. The subsampling convolutions, the attention and convolution sublayers within each block, and the CTC decoder are unchanged. Two MoE configurations are used: an English configuration with  $N = 4$  experts per layer and a multilingual configuration with  $N = 8$  experts per layer.

#### MoE feed-forward layer

In the dense baseline, each Macaron FFN sublayer is a two-layer network: a linear expansion from  $d = 512$  to  $4d = 2048$ , a GELU activation, and a linear projection back to  $d$ . In the MoE variant, this single FFN is replaced by  $N$  independent expert networks. Each expert has the same architecture as the dense FFN, with its own separate parameters. For a given input token, the MoE layer output is the routing-weighted sum of the selected experts' outputs, as described in Section 2.2.2.

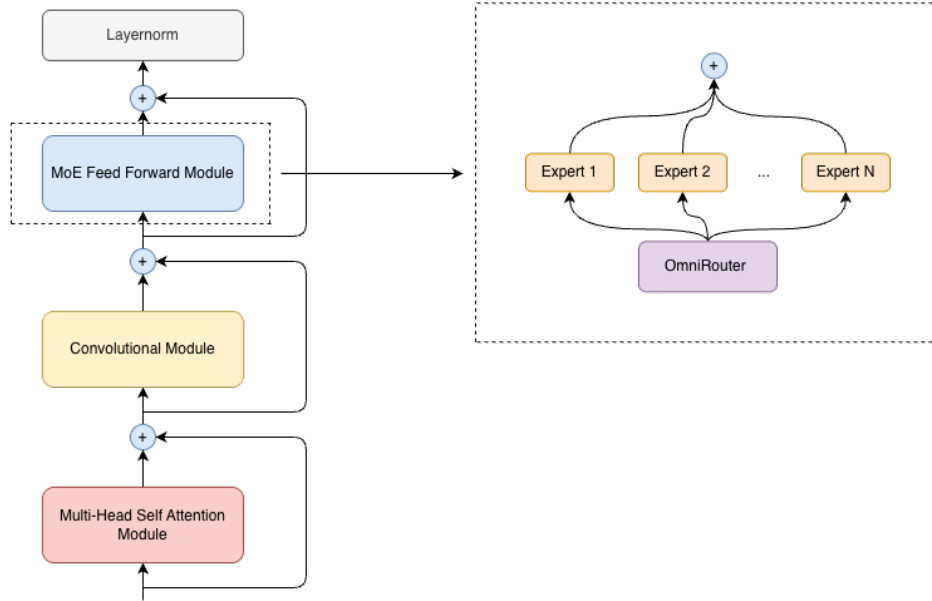


**Figure 3.1:** Left: a single Conformer block in the MoE model. The second feed-forward sublayer is replaced by the MoE Feed Forward Module; the multi-head self-attention and convolution sublayers are unchanged from the dense baseline. Right: inside the MoE Feed Forward Module. The OmniRouter scores all  $N$  experts, selects the top-2, and their outputs are combined in a weighted sum. Because only 2 of the  $N$  experts are active per token, the computation per forward pass matches the dense baseline regardless of how many experts the model contains.

#### Token routing

Figure 3.1 shows the structure of the modified Conformer block and the MoE feed-forward module. Routing follows the sparse top- $k$  formulation described in Section 2.2.2, with  $k = 2$ : each input token is processed by exactly two of the  $N$  experts, regardless of the total number of experts in the layer.

Routing is handled by an OmniRouter [29] with a single set of shared parameters used across all 17 MoE layers (`use_shared_router: true`). The rationale for the shared design is described in Section 2.2.2.

With  $d = 512$  and  $N$  experts, the router adds only  $N \times 512$  parameters in total, several orders of magnitude fewer than the expert FFN parameters.

The router has no access to metadata about speaker age, language, or recording condition. Expert assignments are determined solely from the learned hidden representation of each token, so any age- or language-related specialisation must emerge implicitly from the acoustic signal.

#### Load balancing

Expert collapse and mitigation strategies are described in Section 2.2.2. Two mechanisms are used here. First, Gaussian noise is added to the router logits before top- $k$  selection during training, following noisy top- $k$  gating [45]. In this implementation, the noise standard deviation is set to  $\sigma = 0.005$ . Second, an auxiliary load-balancing loss is added to the CTC objective to encourage more balanced expert utilisation [44]. Its weight is set to  $\lambda = 0.002$  during adult-speech fine-tuning.

During child fine-tuning, two variants are compared. Load balancing encourages the router to distribute tokens more evenly across experts; disabling it allows stronger expert specialisation, but increases the risk that only a small subset of experts is used.

- **LB-on**: load balancing remains active ( $\lambda = 0.002$ ,  $\sigma = 0.005$ ).
- **LB-off**: load balancing is disabled ( $\lambda = 0$ ,  $\sigma = 0$ ).

#### 3.5.3. Parameter counts

Table 3.4 summarises the parameter counts for the three model configurations. In both MoE configurations, only  $k = 2$  of the  $N$  experts are active per token per block, so the active parameter count at inference matches the dense baseline.

**Table 3.4:** Parameter counts for each model configuration. Active parameters are those evaluated in a single forward pass. Counts exclude the tokenizer embedding and CTC output layer.

| Configuration              | Experts | Total params          | Active params         |
|----------------------------|---------|-----------------------|-----------------------|
| Dense FastConformer-CTC    | —       | $\approx 110\text{M}$ | $\approx 110\text{M}$ |
| MoE FastConformer-CTC (EN) | 4       | $\approx 180\text{M}$ | $\approx 110\text{M}$ |
| MoE FastConformer-CTC (ML) | 8       | $\approx 323\text{M}$ | $\approx 110\text{M}$ |

## 3.6. Experimental Setup

### 3.6.1. Compared model configurations

Table 3.5 lists all model configurations evaluated in this work. Model names use an EN- prefix for Experiment 1 models (English only) and an ML- prefix for Experiment 2 models (multilingual); the prefix appears throughout the thesis so that every model reference is unambiguous without needing to consult the table again. Within each experiment, two MoE child fine-tuning variants are compared: one with load balancing disabled and one with it enabled. In Experiment 1 these are **EN-MoE<sub>LB-off</sub>** and **EN-MoE<sub>LB-on</sub>**; in Experiment 2 they are **ML-MoE<sub>LB-off</sub>** and **ML-MoE<sub>LB-on</sub>**. This pairing isolates the effect of load-balancing regularisation on child speech adaptation.

### 3.6.2. Two-stage training protocol

All models share the same two-stage training protocol, illustrated in Figure 3.2. The dense FastConformer-CTC baseline and the MoE encoder are both initialised from a publicly available NVIDIA FastConformer-CTC Large checkpoint<sup>1</sup>. Feed-forward weights are randomly re-initialised before any fine-tuning (Section 3.7.1); all other encoder weights are transferred directly from the checkpoint.

**Stage 1: Adult speech fine-tuning.** Because the feed-forward weights were just randomly re-initialised, and, for the MoE encoder, the router and experts start from the equal-weight initialisation described in Section 3.7.1, neither model yet reproduces the checkpoint’s original adult-speech performance. Stage 1 re-establishes this: the dense baseline and the MoE encoder are both fine-tuned on adult

<sup>1</sup>At the time of writing, the checkpoint is available at [https://huggingface.co/nvidia/stt\\_en\\_fastconformer\\_ctc\\_large](https://huggingface.co/nvidia/stt_en_fastconformer_ctc_large).

**Table 3.5:** All model configurations evaluated in this thesis. EN- prefix: Experiment 1 (English only); ML- prefix: Experiment 2 (multilingual). Subscripts denote the fine-tuning data: LS = LibriSpeech and CV = Common Voice (Stage 1 adult-only fine-tuning); MyST and child = the Stage 2 adult+child mixture (child = combined mixture across all four languages). *Experts* = MoE experts per encoder layer; LB = load-balancing weight  $\lambda$  during training; dash = not applicable.

| Model                           | Encoder | Experts | LB ( $\lambda$ ) | Exp. |
|---------------------------------|---------|---------|------------------|------|
| <b>EN-Dense<sub>LS</sub></b>    | Dense   | —       | —                | 1    |
| <b>EN-Dense<sub>MyST</sub></b>  | Dense   | —       | —                | 1    |
| <b>EN-MoE<sub>LS</sub></b>      | MoE     | 4       | 0.002            | 1    |
| <b>EN-MoE<sub>LB-off</sub></b>  | MoE     | 4       | 0                | 1    |
| <b>EN-MoE<sub>LB-on</sub></b>   | MoE     | 4       | 0.002            | 1    |
| <b>ML-Dense<sub>CV</sub></b>    | Dense   | —       | —                | 2    |
| <b>ML-Dense<sub>child</sub></b> | Dense   | —       | —                | 2    |
| <b>ML-MoE<sub>CV</sub></b>      | MoE     | 8       | 0.002            | 2    |
| <b>ML-MoE<sub>LB-off</sub></b>  | MoE     | 8       | 0                | 2    |
| <b>ML-MoE<sub>LB-on</sub></b>   | MoE     | 8       | 0.002            | 2    |

speech only. Experiment 1 (English only) uses LibriSpeech; Experiment 2 (multilingual) uses Common Voice for Dutch, German, and Polish and LibriSpeech for English (corpus sizes in Table 3.1). In Experiment 2, each language is assigned a fixed 25% share of every training batch regardless of corpus size; this equal language probability sampling prevents larger corpora from dominating training. Stage 1 produces four adult-only models: **EN-Dense<sub>LS</sub>** and **EN-MoE<sub>LS</sub>** for Experiment 1, and **ML-Dense<sub>CV</sub>** and **ML-MoE<sub>CV</sub>** for Experiment 2.

**Stage 2: Child-speech fine-tuning.** Starting from the Stage 1 checkpoints, both models are adapted on a mixture of child and adult speech. The adult data from Stage 1 (LibriSpeech for Experiment 1 (English), LibriSpeech plus Common Voice for Experiment 2 (multilingual)) is retained alongside the child data to prevent catastrophic forgetting of adult-speech performance. In Experiment 1, MyST and LibriSpeech are mixed; in Experiment 2, the 25%-per-language sampling is preserved and split equally between each language’s child corpus (MyST, JASMIN, KidsTALC, PAVSig) and adult corpus. Stage 2 produces the remaining six models: **EN-Dense<sub>MyST</sub>**, **EN-MoE<sub>LB-off</sub>**, and **EN-MoE<sub>LB-on</sub>** for Experiment 1, and **ML-Dense<sub>child</sub>**, **ML-MoE<sub>LB-off</sub>**, and **ML-MoE<sub>LB-on</sub>** for Experiment 2. Exact per-subset weights, step counts, and load-balancing settings are given in Section 3.7.

## 3.7. Training Setup

### 3.7.1. Checkpoint initialisation

Both encoder architectures are initialised from the publicly available NVIDIA NeMo checkpoint mentioned in Section 3.6.2, which was trained using the NeMo toolkit<sup>2</sup> [58]. Starting from a pre-trained checkpoint rather than random initialisation serves two purposes. First, the encoder already contains general acoustic representations (phoneme boundaries, prosodic contours, speaker-level variation) accumulated over 15,000 hours of diverse English speech. These representations transfer to both fine-tuning stages and reduce the amount of labelled data required to reach convergence, which matters given the limited size of child speech corpora. Second, initialising both the dense and MoE architectures from the same checkpoint controls for pre-training history: any difference in downstream performance reflects the architecture rather than a difference in what the encoder has seen before fine-tuning.

This checkpoint was trained on NeMo ASRSet 3.0, a composite English corpus of approximately 15,000 hours comprising LibriSpeech (960 h), Fisher Corpus, Switchboard-1 Dataset, WSJ-0 and WSJ-1, National Speech Dataset (Parts 1 and 6), VCTK, VoxPopuli (EN), Europarl-ASR (EN), Multilingual LibriSpeech EN (2,000-h subset), Mozilla CommonVoice v7.0, and People’s Speech (12,000-h subset). The checkpoint uses the same decoder architecture as this thesis, a single linear CTC output projection. The projection weights themselves are not transferred, however, because their shape is tied to

<sup>2</sup>The code is available at <https://github.com/a-sasin/NeMo>, which is a fork of the original repository available at <https://github.com/NVIDIA-NeMo>.

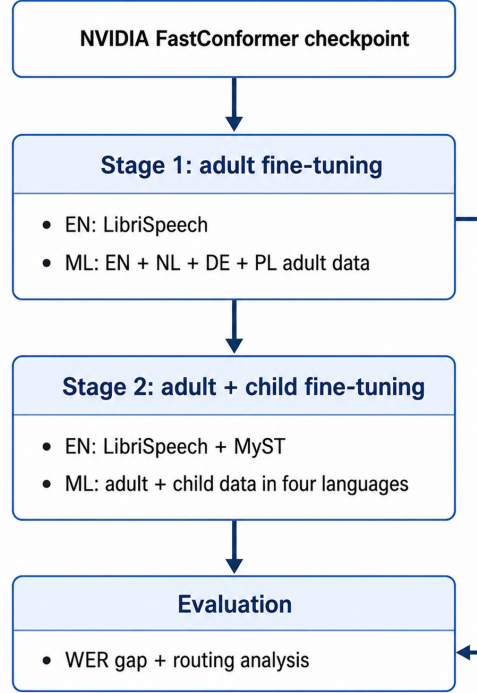


**Figure 3.2:** Two-stage training protocol used in both experiments. Both encoder architectures follow the same stages. Stage 1 fine-tunes on adult speech only, and the resulting adult-only models are also evaluated; Stage 2 adapts from the Stage 1 checkpoint on a mixture of adult and child speech. EN = English experiment; ML = multilingual experiment.

the checkpoint’s original vocabulary, which differs from the vocabularies of the tokenizers trained in Section 3.4. During conversion, the output projection is instead resized to the new vocabulary and randomly initialised. This reinitialisation is identical for the dense and MoE models, keeping the encoder architecture as the only difference between the compared models.

To keep WER comparisons between the dense and MoE models attributable to the architecture rather than to differences in pre-training history, the feed-forward weights inside each encoder block, distinct from the CTC output decoder discussed above, must be reset to the same starting point for both models. Attention and convolutional layers remain unchanged, as they encode general acoustic pattern detectors that are useful regardless of language or speaker type. The feed-forward sublayers, by contrast, encode the most task-specific transformations and are the component being replaced or extended by MoE; resetting them means that any performance difference in RQ1 and RQ2 reflects the architecture alone. Concretely: before Stage 1 training, the feed-forward projection weights in every encoder block are discarded and randomly re-initialised with Kaiming uniform values [59], with bias terms set to zero. All other weights (self-attention, depthwise convolution, and layer normalisation) are transferred directly from the checkpoint.

For the MoE encoder, a further goal is to give the router an equal starting point across all experts so that any expert specialisation observed in RQ3 emerges from the training data rather than from an accidental advantage at initialisation. The NVIDIA checkpoint has no MoE structure and cannot be loaded directly into a model with  $N$  expert networks, so the conversion proceeds in two steps. First, the first feed-forward sublayer in each Conformer block is removed to match the architecture in Figure 3.1, and the remaining feed-forward weights are reset to Kaiming uniform as above. Second, those reset weights are copied identically into each of the  $N$  expert networks, so every expert starts with the same parameters. The router is initialised separately with small random values. Starting all experts from identical weights means the router has no basis to prefer one expert over another at the beginning of training; specialisation emerges gradually as different experts are pushed toward different parts of the input distribution. Separate conversions are performed for the English ( $N = 4$ ) and multilingual ( $N = 8$ ) configurations.

### 3.7.2. Optimisation settings

Both the dense and MoE model variants were optimised using the AdamW optimiser [60], the decoupled-weight-decay variant of Adam [61], with momentum coefficients  $\beta_1 = 0.9$  and  $\beta_2 = 0.98$ , and weight decay  $10^{-3}$ . A cosine annealing learning rate schedule was applied throughout all training stages, with a linear warmup over the first 15,000 steps and a minimum learning rate of  $10^{-4}$ .

The initial learning rate differed between architectures: the dense FastConformer used  $lr = 10^{-3}$ , while the MoE variant used  $lr = 3 \times 10^{-4}$  at all stages. This follows established practice in MoE training, where sparse routing introduces additional gradient noise and can destabilise training at learning rates suited to dense architectures, motivating a lower learning rate for the MoE variant [44]. Fixing a single shared learning rate for both architectures would not have made the comparison more controlled: it would instead risk undertraining whichever architecture that value does not suit, confounding the comparison in the opposite direction. Gradient clipping was applied with a maximum norm of 1.0 during fine-tuning stages. All models were trained in bfloat16 mixed precision.

#### Batching and hardware

All experiments were conducted on a single node with four NVIDIA A100 80 GB GPUs using the Distributed Data Parallel (DDP) strategy implemented in the NVIDIA NeMo Toolkit [58].

To maximise GPU utilisation while avoiding out-of-memory errors, Lhotse’s dynamic bucketing was used with per-bucket batch sizes determined by the OOMptimizer algorithm [49]. Duration bins (30 bins) were estimated from the training data distribution using NeMo’s `estimate_duration_bins.py`. For each duration bucket, OOMptimizer creates temporary dummy batches whose tensor shapes correspond to the maximum sequence length in that bucket. These dummy batches are used only to run memory stress tests: OOMptimizer performs a bisection search over batch sizes and identifies the largest batch size that fits in GPU memory. The dummy batches are generated on the fly for profiling and are discarded afterwards; they are never used for model optimisation. Because the MoE model has higher memory requirements than the dense FastConformer (due to the additional expert parameters), batch sizes were profiled using the MoE configuration. The resulting per-bucket batch-size schedule was then applied to both the MoE and dense models. This ensures that both architectures are trained with the same bucketing scheme and batch-size schedule, so WER comparisons are controlled for data exposure rather than confounded by different effective batch sizes.

### 3.7.3. Stage 1: Adult speech fine-tuning

In Experiment 1, **EN-Dense<sub>LS</sub>** and **EN-MoE<sub>LS</sub>** are fine-tuned on the full 960-hour LibriSpeech training set. In Experiment 2, **ML-Dense<sub>CV</sub>** and **ML-MoE<sub>CV</sub>** are fine-tuned on the four-language adult corpus (Table 3.1), with each language assigned a fixed 25% share of every training batch regardless of corpus size; this equal language sampling prevents the larger corpora from dominating training. Training ran for up to 100,000 steps for Experiment 1 and 200,000 steps for Experiment 2, with validation every 5,000 steps; the checkpoint with the lowest validation WER is retained. The MoE encoder uses the load-balancing auxiliary loss ( $\lambda = 0.002$ ,  $\sigma = 0.005$ ) throughout Stage 1.

### 3.7.4. Stage 2: Child-speech fine-tuning

Stage 2 adapts the Stage 1 models to child speech, producing the models evaluated against RQ1 and RQ2. Fine-tuning on child speech alone was tested first, but this caused the models to overfit to child speech acoustics and substantially degrade adult WER. Adult speech is therefore retained in the Stage 2 mixture alongside child data. This is necessary because RQ1 and RQ2 measure the adult-child WER *gap*: a model that improves child WER while collapsing adult WER could still show a smaller gap, but only by levelling adult performance down to meet child performance, not by making the model genuinely better at recognising child speech. Reporting adult WER alongside the gap, as described in Section 3.8.2, is what exposes this failure mode.

For Experiment 1, fine-tuning starts from **EN-Dense<sub>LS</sub>** and **EN-MoE<sub>LS</sub>**, with MyST added to the LibriSpeech data. The sampling weights are: MyST at 0.50, `train-other-500` at 0.20, `train-clean-360` at 0.15, and `train-clean-100` at 0.15. This produces three Stage 2 Experiment 1 models, each trained for up to 100,000 steps. Alongside the dense model (**EN-Dense<sub>MyST</sub>**), two MoE configurations are trained: **EN-MoE<sub>LB-off</sub>**, in which load balancing is disabled ( $\lambda = 0$ ,  $\sigma = 0$ ) so routing can evolve freely

during child adaptation, and **EN-MoE<sub>LB-on</sub>**, in which load balancing stays active as in Stage 1. Comparing these two models tests whether enforcing uniform expert utilisation helps or hinders child-speech adaptation.

For Experiment 2, fine-tuning starts from **ML-Dense<sub>cv</sub>** and **ML-MoE<sub>cv</sub>**. The 25% language share from Stage 1 is retained, and within each language the share is split equally between the child corpus (MyST, JASMIN, KidsTALC, or PAVSig) and the adult corpus. The model is therefore adapted to paired adult and child sources for all four languages simultaneously, while high-resource adult corpora are prevented from dominating the adaptation stage. All three Stage 2 multilingual models (**ML-Dense<sub>child</sub>**, **ML-MoE<sub>LB-off</sub>**, and **ML-MoE<sub>LB-on</sub>**) train for up to 100,000 steps, with the same load-balancing settings as their English counterparts.

## 3.8. Evaluation Metrics

### 3.8.1. Word error rate

The primary evaluation metric is Word Error Rate (WER), defined as

$$\text{WER} = \frac{S + D + I}{N} \times 100\%, \quad (3.1)$$

where  $S$ ,  $D$ , and  $I$  are the number of substitution, deletion, and insertion errors, respectively, and  $N$  is the total number of reference words. WER is computed using greedy CTC decoding without a language model: at each frame, the token with the highest predicted probability is taken directly, and the collapsing rule from Section 2.1.1 is applied to obtain the transcript, with no beam search over alternative hypotheses and no external language model rescoring. This matches the decoding strategy used during training and validation.

### 3.8.2. Performance disparity metric

The main disparity metric is the adult-child WER gap, equivalent to the group-to-norm bias difference with adult speech as the reference group and child speech as the evaluated group [62]:

$$\text{Gap}_{\text{child-adult}} = G2norm_{\text{diff}}^{\text{child}} = \text{WER}_{\text{child}} - \text{WER}_{\text{adult}}. \quad (3.2)$$

WER values are expressed in percent, so the gap is in percentage points (pp). Positive values indicate higher recognition error for child speech than for adult speech; values closer to zero indicate a smaller disparity. The gap is reported alongside adult and child WER because a lower child WER alone is not sufficient: a model should reduce the disparity without substantially degrading adult-speech performance.

### 3.8.3. Statistical significance testing

Because the compared systems are evaluated on the same test utterances, WER differences are paired rather than independent. Statistical significance is tested using paired speaker-level non-parametric bootstrap resampling, following ASR significance-testing practice [63], [64]. The resampling unit is the speaker whenever speaker identity is available: speakers are sampled with replacement, and all utterances from each sampled speaker are included in the bootstrap sample. This preserves within-speaker dependence and avoids treating multiple utterances from the same speaker as independent observations.

Two types of comparisons are tested. First, when comparing two models on the same adult or child test set, the tested statistic is the WER difference:

$$\Delta \text{WER} = \text{WER}_A - \text{WER}_B. \quad (3.3)$$

Negative values mean that Model A has a lower WER than Model B on a particular test set.

Second, when testing whether one model reduces the adult-child gap more than another, the tested statistic is the difference between their gaps:

$$\Delta \text{Gap} = (\text{WER}_{A,\text{child}} - \text{WER}_{A,\text{adult}}) - (\text{WER}_{B,\text{child}} - \text{WER}_{B,\text{adult}}). \quad (3.4)$$

Negative values mean that Model A has a smaller adult-child gap than Model B. Thus,  $\Delta\text{WER}$  captures performance differences on a single test set, while  $\Delta\text{Gap}$  captures whether the adult-child disparity itself changes. Adult and child speaker pools are resampled independently because the adult and child evaluation sets contain disjoint speakers.

For each comparison, 10,000 bootstrap samples are drawn; each draw resamples a different set of speakers, so WER must be calculated fresh for every sample rather than reused from the original test-set evaluation. For a given bootstrap sample, the substitution, deletion, insertion, and reference-word counts (Equation 3.1) are first summed across all utterances of the sampled speakers, and WER is then computed from these summed counts. This keeps WER a ratio of pooled error counts to pooled reference words, matching its corpus-level definition, rather than an average of per-speaker WER values, which would bias the estimate since WER is not a linear function of utterance-level counts. For each resample, the relevant tested statistic ( $\Delta\text{WER}$  or  $\Delta\text{Gap}$ ) is computed. The 2.5th and 97.5th percentiles of its bootstrap distribution define the 95% confidence interval for the effect size.

Bootstrap  $p$ -values are computed from the resampled distribution of the tested statistic. For a two-sided test, the bootstrap  $p$ -value is estimated from the proportion of resampled statistics that fall on the opposite side of zero from the observed effect, doubled to account for both tails. Since each experiment contains multiple related comparisons, these bootstrap  $p$ -values are corrected using the Benjamini-Hochberg false discovery rate procedure at  $q = 0.05$  [65]. A comparison is treated as statistically significant when the adjusted  $p$ -value is below 0.05.

## 3.9. Experiments

### 3.9.1. Experiment 1: English adult-child WER gap

**RQ1:** Does integrating sparse Mixture-of-Experts layers into a FastConformer-CTC encoder reduce the Word Error Rate (WER) gap between adult and child speakers compared to a dense FastConformer baseline, while not degrading adult-speaker performance?

Experiment 1 trains five English models in total (Table 3.5). Stage 1 produces the adult-only baselines **EN-Dense<sub>LS</sub>** and **EN-MoE<sub>LS</sub>**. Stage 2 adapts these to child speech, producing three models: the dense baseline **EN-Dense<sub>MyST</sub>** and two MoE variants, **EN-MoE<sub>LB-off</sub>** and **EN-MoE<sub>LB-on</sub>**, which differ only in whether load balancing is applied during Stage 2 (Section 3.7.4). The Stage 1 models test whether MoE generalises better to child speech from adult-only fine-tuning, before any child speech audio enters training. Stage 2 tests whether that advantage is additive with child-speech fine-tuning.

All five models are evaluated on LibriSpeech test set and the MyST test set (child). The adult-child gap (Equation 3.2) is compared at both stages: a smaller value for the MoE models indicates a reduced adult-child disparity. Adult WER is reported alongside to verify that child-speech adaptation does not cause forgetting of adult performance.

### 3.9.2. Experiment 2: Multilingual generalisation

**RQ2:** When trained on multilingual adult and child speech in English, Dutch, German, and Polish, how does the MoE encoder affect the adult-child WER gap relative to a dense baseline, and does this effect vary across languages?

Experiment 2 trains five multilingual models in total (Table 3.5), following the same two-stage procedure as Experiment 1 but across four languages and using 8 experts in the MoE encoder. Stage 1 produces the adult-only baselines **ML-Dense<sub>CV</sub>** and **ML-MoE<sub>CV</sub>**. Stage 2 produces three child-adapted models: **ML-Dense<sub>child</sub>**, **ML-MoE<sub>LB-off</sub>**, and **ML-MoE<sub>LB-on</sub>**.

WER and the adult-child gap are reported per language. Adult test sets are the per-language Common Voice test partitions and the combined LibriSpeech test set. Child test sets are JASMIN (Dutch), KidsTALC (German), PAVSig (Polish), and MyST (English).

### 3.9.3. Experiment 3: Expert routing analysis

**RQ3:** Without explicit age labels, do routing distributions in a trained MoE encoder differ between adult and child speech within the same language, and do the experts that receive age-differentiated routing compute functionally distinct representations?

No additional training is needed. Six MoE models are analysed in total: **EN-MoE<sub>LS</sub>**, **EN-MoE<sub>LB-off</sub>**, and **EN-MoE<sub>LB-on</sub>** for the English analysis, and **ML-MoE<sub>CV</sub>**, **ML-MoE<sub>LB-off</sub>**, and **ML-MoE<sub>LB-on</sub>** for the multilingual analysis. Each model is run in evaluation mode, meaning weights are fixed (no gradient updates), and stochastic training-only operations such as SpecAugment and router noise are disabled, so routing depends only on the input and the trained parameters. Expert assignments are recorded at all 17 encoder layers for every test utterance. Four analyses are then conducted:

1. **English age routing.** Routing distributions of adult (LibriSpeech) and child (MyST) speech are compared at each layer using Jensen-Shannon Divergence (JSD), a symmetric, bounded measure of the difference between two probability distributions; JSD is defined and motivated in Section 3.10.2. **EN-MoE<sub>LS</sub>** is included as a pre-child-exposure routing baseline.
2. **Functional divergence.** At the layers with the highest adult-child JSD (the two peak regions identified in the routing analysis), all four experts are applied to the same input frames with the router bypassed. Pairwise cosine similarity between expert outputs measures whether differently-routed experts actually compute different transformations.
3. **Domain control.** A YODAS2 subset (adult YouTube speech, not in training data) is routed through the same models. If routing differences between LibriSpeech and MyST reflect speaker age rather than recording conditions, YODAS2 (a different adult domain) should route similarly to LibriSpeech and differently from MyST.
4. **Multilingual routing.** Using the ML-MoE models, routing is first compared across the four languages on adult data only, to measure implicit language specialisation. It is then compared between adult and child within each individual language, holding language constant so that any remaining routing difference is consistent with age-related acoustics, although corpus and elicitation differences between adult and child recordings remain possible contributors.

The extraction procedure and metrics are described in Section 3.10.2.

## 3.10. Routing Analysis

The WER results reported in Chapter 4 show whether MoE reduces the adult-child performance gap, but they do not reveal the underlying mechanism. Two mechanisms could explain such a reduction. First, the MoE model has more total parameters than the dense baseline; additional capacity alone could improve performance on any speaker group, with no routing specialisation involved. Second, the router may learn to send adult and child speech to different experts, so that each expert processes a more homogeneous acoustic distribution and becomes better calibrated for its assigned speech type. These two mechanisms are not mutually exclusive, but they have different implications: capacity alone does not require the router to do anything meaningful, whereas routing specialisation implies that the model has implicitly learned to distinguish adult from child speech without age labels.

### 3.10.1. Analysis overview

The routing analysis is organised around two questions. First, in the English setting, it tests whether child-speech fine-tuning changes how adult and child speech are assigned to experts. This is examined through adult-child routing divergence, expert preference patterns, a router-bypass functional analysis, and a YODAS2 domain-control experiment. The goal is to determine whether any observed routing differences are consistent with age-related acoustics rather than only with recording-domain differences.

Second, in the multilingual setting, the analysis tests how age-related routing interacts with language-related routing. Adult speech from English, Dutch, German, and Polish is first compared to determine whether the router implicitly specialises by language. Adult and child speech are then compared within each language to test whether age-related routing remains visible when language is held constant.

Because these analyses use the same extraction procedure and metrics, the shared routing methodology is described before the experiment-specific analyses.

### 3.10.2. Shared routing extraction and metrics

#### Routing extraction

Each model is run in evaluation mode on its respective test sets, and the expert indices assigned by the router are recorded at each of the 17 MoE layers for every utterance. Each encoder block contains one MoE layer, giving 17 layers in total. With top- $k = 2$  routing, every input frame is assigned to two experts following standard sparse MoE routing [44], [45]; these assignments are aggregated per utterance to produce a per-utterance expert usage distribution. The result is one routing distribution per group (e.g., adult English, child English) per layer.

#### Metrics

The routing-analysis methodology builds on standard sparse-MoE routing, where a gating network selects a small subset of experts for each input [44], [45]. The comparison of routing distributions across groups is adapted from Bandarkar et al. [66], who analyse multilingual MoE routing across languages. In this thesis, this idea is transferred to ASR by aggregating routing over acoustic frames and comparing adult and child speaker groups. The duration-matched bootstrap, router-bypass functional expert analysis, and YODAS2 domain-control experiment are thesis-specific extensions for the adult-child ASR setting.

The following metrics are all computed per layer and per group. A group is a set of utterances sharing the same language and speaker type (e.g., adult English or child Dutch). Let  $N$  represent the number of experts per layer (4 in Experiment 1, 8 in Experiment 2), and let  $k = 2$  be the number of experts selected per frame by the router.

**Routing distribution.** For a group  $g$  at a given layer, let  $F^{(g)}$  be the total number of input frames across all utterances in that group, and let  $c_e^{(g)}$  be the number of times expert  $e$  is selected across those frames. Because each frame is assigned to  $k$  experts, the total number of assignments is  $k \cdot F^{(g)}$ . The routing distribution of group  $g$  is the vector

$$p_e^{(g)} = \frac{c_e^{(g)}}{k \cdot F^{(g)}}, \quad e = 1, \dots, N, \quad (3.5)$$

which sums to 1 over all experts and can be read as the probability that a randomly sampled expert assignment from group  $g$  at this layer belongs to expert  $e$ . This count-based distribution is used as a summary of sparse top- $k$  expert assignment [44], [45].

**Expert load.** The expert load measures how much of the input a given expert handles, relative to what it would handle under perfectly uniform routing. It is defined as

$$\ell_e^{(g)} = \frac{c_e^{(g)}}{F^{(g)}} = k \cdot p_e^{(g)}, \quad (3.6)$$

which is the average number of expert assignments received by expert  $e$  per input frame in group  $g$ . With uniform top- $k$  routing, each expert would receive  $\ell_e^{(g)} = k/N$ . Values above this baseline indicate experts that are used more than average for that group (adult or child speakers), while values below it indicate underused experts. This follows the standard MoE interpretation of expert utilisation and load balancing [44], [45].

**Expert preference delta.** To identify which experts are preferentially activated by adult or child speech, the load difference between the two groups is computed as

$$\Delta_e = \ell_e^{(\text{child})} - \ell_e^{(\text{adult})}, \quad (3.7)$$

following the approach of Bandarkar et al. [66]. A positive value indicates that expert  $e$  is selected more often for child speech than for adult speech; a negative value indicates adult preference. The uniform-routing value  $k/N$  is used only as an interpretive reference for the load profiles, not as the

direct baseline in  $\Delta_e$ . This is reported as a heatmap across all experts and layers, showing where and how expert preferences differ between age groups.

**Jensen-Shannon Divergence.** To summarise how different the routing distributions of two groups are at a given layer, this thesis uses the Jensen-Shannon Divergence (JSD) [67], a symmetric measure of the difference between two probability distributions, based on the Kullback-Leibler (KL) divergence [68]:

$$\text{KL}(P \parallel Q) = \sum_{e=1}^N p_e \log_2 \frac{p_e}{q_e}. \quad (3.8)$$

KL is asymmetric ( $\text{KL}(P \parallel Q) \neq \text{KL}(Q \parallel P)$  in general); JSD symmetrises it by comparing both distributions to their average  $M = (P + Q)/2$ :

$$\text{JSD}(P \parallel Q) = \frac{1}{2} \text{KL}(P \parallel M) + \frac{1}{2} \text{KL}(Q \parallel M). \quad (3.9)$$

With the  $\log_2$  base, JSD is bounded in  $[0, 1]$ : 0 means the two routing distributions are identical (adult and child frames sent to the same experts in the same proportions), and 1 means they share no overlap at all (the groups are always routed to completely different experts). In practice, values are expected to be small, since adult and child speech share the bulk of their phonetic content and the load-balancing loss pushes expert utilisation toward uniformity; even a modest but consistent JSD across several layers is therefore informative.

JSD is computed at the group level: one routing distribution is estimated per group (e.g. all adult English utterances, all child English utterances) at each layer, and JSD is computed between the two group distributions. This adapts the approach of Bandarkar et al. [66], who isolate language as the single varying factor by comparing routing on content-matched sentence pairs across languages. ASR data has no equivalent, since no corpus provides the same utterance spoken by both an adult and a child, so distributions are instead compared in aggregate, pooling routing statistics over all utterances within each group.

Bandarkar et al. [66] also normalise JSD by mixture entropy, to correct for layers where routing concentrates on one or two experts and both compared distributions become artificially similar regardless of any real difference. Routing entropy in the present model stays near-maximal at every layer (1.86-2.00 of a possible 2.00 bits for 4 experts), so this correction would not change the relative pattern; raw JSD is reported instead.

JSD is computed separately at each of the 17 layers, producing a per-layer profile that shows where in the encoder routing differences are largest. Bootstrap 95% confidence intervals are estimated over 1,000 duration-matched resamples, where the larger group is randomly downsampled to match the total duration of the smaller group so that both groups contribute equal statistical weight. Fewer resamples are used than for the WER significance tests (Section 3.8.3) because each JSD resample recomputes routing distributions across all 17 layers, which is substantially more expensive than pooling error counts.

**Bootstrap confidence intervals.** The JSD value at each layer is estimated from the test utterances actually available (LibriSpeech `test-clean` and the MyST test partition), a finite sample rather than every utterance any adult or child could ever produce. If a different, similarly sized sample of adult and child speech had been used instead, the measured JSD could come out slightly higher or lower purely because of which utterances happened to be included, even if the model's underlying routing behaviour were exactly the same. To estimate how much the measured JSD could vary for this reason, bootstrap resampling is used [64].

The procedure works as follows. First, the two groups - adult and child speakers - are duration-matched: the group with more total audio is randomly downsampled until its total duration matches the smaller group. This ensures that both groups contribute the same amount of speech to the comparison. Second, utterances are randomly drawn with replacement from each group, keeping the original group sizes. Some utterances may be drawn more than once, and others not at all. JSD is then computed

from these resampled groups. This process is repeated 1,000 times, producing 1,000 JSD values. The 2.5th and 97.5th percentiles of these values define the 95% confidence interval.

The confidence interval is used to assess the stability of the estimated routing divergence. A narrow interval indicates that the estimated JSD is robust to the sampled utterances, whereas a wide interval indicates greater sampling variability.

### 3.10.3. English models routing analysis

Three English MoE models are analysed: **EN-MoE<sub>LS</sub>**, **EN-MoE<sub>LB-off</sub>**, and **EN-MoE<sub>LB-on</sub>**. **EN-MoE<sub>LS</sub>** is fine-tuned only on adult LibriSpeech and serves as the pre-child-exposure routing baseline. Comparing this model with the two child-adapted models tests whether child fine-tuning changes adult-child routing behaviour. Comparing **EN-MoE<sub>LB-off</sub>** with **EN-MoE<sub>LB-on</sub>** tests whether load-balancing regularisation affects the resulting routing patterns.

The English analysis follows a two-step interpretation logic. First, routing divergence is used to test whether adult and child speech are assigned to different experts after child-speech fine-tuning. Second, the router-bypass functional analysis tests whether the most age-specialised experts also compute different transformations. Routing divergence alone does not show that this specialisation has any computational consequence: if adult and child speech were routed to different experts, but those experts computed near-identical outputs (Section 3.10.3), the divergence would not reflect a real difference in what the model does with the two speech types. Age-specialised routing is therefore interpreted as evidence of genuine functional specialisation only where routing differences and functional differences occur in the same layers and expert pairs; agreement between the two analyses is what distinguishes this from a difference in expert index alone.

#### Age routing analysis

Adult and child routing distributions are compared within English. LibriSpeech `test-clean` serves as the adult group, and the MyST test partition as the child group. JSD, expert load, and expert preference delta are computed for each model across all 17 layers using the shared procedure in Section 3.10.2.

#### Functional expert analysis

The functional analysis tests the second part of the interpretation logic: whether differently routed experts also compute different transformations. The analysis focuses on the layers where adult-child JSD is highest, identified in Section 4.3.1. At these layers, the router is bypassed, and each sampled encoder frame is passed through all four experts independently, producing four output representations for the same input frame. Pairwise cosine similarity is then computed between the outputs of all  $\binom{4}{2} = 6$  expert pairs:

$$\text{sim}(E_i, E_j) = \frac{\mathbf{o}_i \cdot \mathbf{o}_j}{\|\mathbf{o}_i\| \|\mathbf{o}_j\|}, \quad (3.10)$$

where  $\mathbf{o}_i$  is the output of expert  $i$  on a given frame. Similarity is averaged across all frames in the evaluation set, giving a  $4 \times 4$  matrix per model per layer. Values close to 1 indicate that two experts produce nearly identical outputs for the same input (functionally interchangeable), whereas lower values indicate genuine functional divergence. Cosine-similarity comparisons of expert outputs have also been used to study functionally similar experts in MoE models [69]; here, cosine similarity is used as a post-hoc probe of whether differently routed ASR experts compute distinct transformations.

The input set uses 200 audio files drawn from the LibriSpeech `test-clean` set, with up to 20,000 frames sampled per layer. Adult speech alone is sufficient here because the analysis compares the outputs of all experts on identical input frames: it measures whether the experts implement different transformations, not how they respond to different speaker groups. A subset is used because the router-bypass analysis is computationally more expensive than routing extraction: each sampled encoder frame must be passed through every expert rather than only through the experts selected by the router. The size of 200 files is therefore a practical compromise between acoustic coverage and computational cost. This subset is used only for the functional diagnostic; the main routing-divergence analysis is computed on the full test sets. Comparing similarity matrices before and after child fine-tuning shows whether functional divergence between experts increases, while comparing the most age-specialised

expert pairs against other pairs tests whether routing specialisation and functional specialisation co-occur at the same layers and experts.

#### Domain control experiment

An alternative explanation for routing differences between child and adult speech is that the router responds to recording conditions rather than age-related acoustics. MyST was recorded in a tutoring environment and has a different noise floor and microphone character than LibriSpeech audiobook recordings. If routing differences were driven by recording-level properties, the finding would not generalise beyond these datasets. This control is motivated by prior child-ASR work showing domain shift when adult-pretrained speech models are adapted to child speech [70].

To test this, routing is analysed on a subset of YODAS2 English (Section 3.2.1). YODAS2 consists of YouTube speech: adult English from a different recording domain than LibriSpeech. Unlike LibriSpeech, YODAS2 was not filtered to exclude child speech, so a small proportion of child utterances may be present; it is treated here as a predominantly adult corpus. YODAS2 was not used for model training. Three groups are compared, as shown in Table 3.6. If routing responds primarily to recording conditions, YODAS2 and LibriSpeech would be expected to route differently, despite both being adult speech. If routing reflects age-related acoustics, the two adult groups should route similarly to each other and differently from the child group.

**Table 3.6:** Groups used in the domain control experiment. All three groups contain English speech. YODAS2 was not used during model training.

| Group            | Age   | Domain         | In training |
|------------------|-------|----------------|-------------|
| LibriSpeech test | Adult | Read audiobook | Yes         |
| MyST test        | Child | Conversational | Yes         |
| YODAS2 subset    | Adult | YouTube speech | No          |

A subset of YODAS2 English at quality bucket 0 is used; JSD and expert load are computed for this subset and compared against LibriSpeech and MyST at each layer following Section 3.10.2.

#### 3.10.4. Multilingual models routing analysis

The multilingual setting introduces language as an additional possible source of routing variation, alongside speaker age. The multilingual routing analysis extends the English routing analysis, using the same routing metrics throughout. Previous work has shown that language-conditioned expert routing can support multilingual ASR, and analyses of multilingual MoE language models have found emergent language-dependent routing patterns [19], [23], [66]. The analysis, therefore, separates language-related routing from age-related routing: without holding language constant, an apparent age-related routing difference could instead reflect the language composition of the speech being compared rather than speaker age itself. Comparing adult and child speech within each language individually (Section 3.10.4) holds language identity constant, so any remaining routing divergence is consistent with age-related acoustic variation, although corpus and elicitation differences between the adult and child corpora remain possible contributors.

Three multilingual MoE models are analysed: **ML-MoE<sub>CV</sub>**, **ML-MoE<sub>LB-off</sub>**, and **ML-MoE<sub>LB-on</sub>**. **ML-MoE<sub>CV</sub>** is fine-tuned only on adult multilingual speech and serves as the adult-only routing baseline. The two child-adapted models are used to test whether child-speech fine-tuning introduces additional age-related routing structure while preserving or modifying any language-related routing structure learned during adult multilingual training.

The multilingual analysis consists of two parts. First, routing is compared across adult speech from English, Dutch, German, and Polish to test whether the router implicitly specialises by language and whether language produces stronger routing separation than age. Second, adult and child routing distributions are compared within each of the four languages to test whether age-related routing remains detectable once language is held constant. Together, these analyses test whether language is the dominant routing factor among the factors examined and whether speaker age provides an additional, weaker source of expert specialisation within language.

### Language specialisation

The multilingual model was trained on four languages simultaneously, but the router receives no explicit language label. Unlike language-routed multilingual ASR systems that explicitly guide or assign experts by language [19], [23], the experts in the present model are not assigned language-specific roles. Any systematic language-dependent expert preferences would therefore constitute emergent specialisation. To test whether such specialisation is present, routing distributions are computed separately for adult English, Dutch, German, and Polish speech. Adult-only datasets are used so that speaker age is held constant across the comparison. JSD is then computed between every pair of adult-language routing distributions at each layer, producing a  $4 \times 4$  matrix of pairwise language distances per layer. Expert load is also reported per language to identify which experts are preferentially activated by each language.

### Age specialisation within language

Knowing that the router distinguishes languages does not tell us whether it also distinguishes adult from child speakers within a language. Because child and adult speech differ systematically in acoustic, articulatory, and pronunciation characteristics [6], [10], speaker age is a plausible additional source of routing variation. The second analysis tests this by repeating the adult-versus-child comparison within each of the four languages: LibriSpeech versus MyST for English, Common Voice Dutch versus JASMIN for Dutch, Common Voice German versus KidsTALC for German, and Common Voice Polish versus PAVSig for Polish. Because each comparison is made within a single language, language identity is held constant, and the remaining routing divergence can be interpreted as age-related acoustic variation, subject to corpus differences between the adult and child datasets.

For each language and model, JSD, expert load, and expert preference delta are computed across all 17 layers using the shared procedure in Section 3.10.2. This tests whether multilingual child adaptation produces age-related routing patterns comparable to those in the English-only model, and whether those patterns are shared across languages or language-specific.

# 4

## Results

This chapter reports WER and adult-child gap results for all experimental conditions. Experiment 1 (Section 4.1) covers the English-only models and examines whether MoE reduces the adult-child WER gap without degrading adult-speaker performance (RQ1). Experiment 2 (Section 4.2) covers the multilingual models and examines whether this effect holds across English, Dutch, German, and Polish (RQ2). Experiment 3 (Section 4.3) reports the routing analysis and examines whether the MoE router develops age-related expert specialisation without explicit age labels (RQ3). Statistical significance in Experiments 1 and 2 is defined as adjusted  $p < 0.05$  after correction for multiple comparisons, using the bootstrap procedure described in Section 3.8.3; each table caption specifies exactly what its superscript markers indicate.

### 4.1. Experiment 1: English Adult-Child WER Gap (RQ1)

After Stage 1 adult-speech fine-tuning, both encoders achieve strong adult WER, but the adult-child gap remains large because neither model has yet been exposed to child speech. In Stage 2, child speech is introduced during fine-tuning, which substantially improves child WER and narrows the gap for both encoders. This improvement is expected and is therefore not treated as evidence for the MoE architecture on its own. The architectural comparison instead concerns whether, within the same training stage, the MoE encoder yields a smaller adult-child gap than the dense baseline. Table 4.1 reports WER and the adult-child gap for all English models at each stage.

After Stage 1 LibriSpeech fine-tuning, the adult-child gap is statistically significantly smaller for the MoE encoder than for the dense model baseline ( $\Delta\text{Gap} = -2.36$  pp, adjusted  $p < 0.001$ ). Although the gap remains large, this provides the first indication that the MoE encoder generalises better to child speech than the dense encoder, despite neither model having been fine-tuned on child speech audio.

After Stage 2 MyST fine-tuning, both models improve substantially on child speech while keeping adult performance nearly unchanged: adult WER shifts by at most 0.10 pp relative to the Stage 1 values for both encoders. Child WER falls by 11.32 pp for the dense model and by approximately 9 pp for both MoE variants. The dense model’s larger absolute gain reflects its worse Stage 1 starting point. Both encoders converge to similar final child WER values, with MoE achieving the best result overall. The adult-child gap is roughly halved for both encoders relative to Stage 1. Both MoE variants reduce the gap significantly beyond the dense baseline:  $\Delta\text{Gap} = -0.36$  pp for EN-MoE<sub>LB-off</sub> (adjusted  $p < 0.001$ ) and  $-0.27$  pp for EN-MoE<sub>LB-on</sub> (adjusted  $p = 0.004$ ).

Disabling load balancing was expected to let the router specialise more freely for child speech, since it removes the pressure to route tokens uniformly across experts; keeping load balancing on was expected to protect adult performance by limiting how far routing drifts toward child speech. The results match this expectation in direction: EN-MoE<sub>LB-off</sub> gives the lower child WER and the smaller gap, while EN-MoE<sub>LB-on</sub> gives the lower adult WER. Neither difference is statistically significant, however, so load balancing has little effect on Stage 2 outcomes: the age-related specialisation found in this thesis is driven mainly by the child fine-tuning data itself, not by the load-balancing loss.

These results support the hypothesis that sparse expert routing lets the encoder generalise to child speech from adult-only fine-tuning, and that this advantage persists after child-speech fine-tuning, answering RQ1 in the affirmative. The Stage 1 result is the more informative finding, because it shows the advantage exists before any child speech audio enters training; child speech is present at this stage only as transcription text in the tokenizer training data (Section 3.4), which affects the subword vocabulary but provides no acoustic exposure. Child speech differs acoustically from the adult speech both encoders were trained on; sparse expert routing lets the MoE encoder route these acoustically distinct inputs to different experts, while the dense encoder processes all speech through the same shared parameters regardless of the difference. Neither model receives an explicit age label at any training stage, so this separation cannot come from age supervision. Child fine-tuning narrows the gap between the two architectures, since both benefit from the added child data. However, the MoE advantage remains statistically significant after fine-tuning: the benefit already visible at Stage 1 does not disappear once child data is introduced, but persists alongside the separate benefit both models gain from the child fine-tuning data itself.

**Table 4.1:** WER results on English test sets at each training stage. Subscripts denote model configuration: LS = LibriSpeech-only fine-tuning (Stage 1, adult speech only); MyST = MyST + LibriSpeech fine-tuning (Stage 2, adult and child speech); LB-on/LB-off = whether the load-balancing auxiliary loss remained active or was disabled during Stage 2 MoE fine-tuning. Gap = Child WER – Adult WER in percentage points. Best value in each column is shown in bold.

| Model                                       | Adult WER (%) | Child WER (%) | Gap (pp)        |
|---------------------------------------------|---------------|---------------|-----------------|
| <i>After LibriSpeech fine-tuning</i>        |               |               |                 |
| EN-Dense <sub>LS</sub>                      | 2.98          | 25.31         | +22.33          |
| EN-MoE <sub>LS</sub>                        | <b>2.56*</b>  | 22.53*        | +19.97*         |
| <i>After MyST + LibriSpeech fine-tuning</i> |               |               |                 |
| EN-Dense <sub>MyST</sub>                    | 2.94          | 13.99         | +11.05          |
| EN-MoE <sub>LB-on</sub>                     | 2.62*         | 13.40*        | +10.78*         |
| EN-MoE <sub>LB-off</sub>                    | 2.66*         | <b>13.35*</b> | + <b>10.69*</b> |

A superscript asterisk (\*) indicates a significant improvement over the dense model at the same training stage for the corresponding metric; for the Gap column, improvement means a smaller adult-child gap.

## 4.2. Experiment 2: Multilingual Generalisation (RQ2)

Experiment 1 shows that the MoE encoder reduces the adult-child WER gap for English. Experiment 2 tests whether this advantage generalises across languages when dense and MoE are trained on all four jointly, rather than as separate per-language models. Both architectures go through stages analogous to Stage 1 and Stage 2 in Experiment 1: adult-only fine-tuning on Common Voice and LibriSpeech, followed by child-adapted fine-tuning with child speech from all four languages. Each resulting model is then evaluated separately per language. Table 4.2 reports adult WER, child WER, and the adult-child gap for all multilingual models, grouped by language; the Gap column is the main evidence for RQ2.

At Stage 1, the multilingual MoE model does not reproduce the English-only Stage 1 advantage: in all four languages, ML-MoE<sub>CV</sub> has a significantly larger adult-child gap than ML-Dense<sub>CV</sub>, driven by worse child WER. This comparison should be read with caution, however, since the multilingual setting is not a direct replication of Experiment 1’s Stage 1: the multilingual MoE router has more experts per layer and receives less training data per language (Section 5.1.2).

At Stage 2, the result becomes language-dependent. English is the only language where MoE significantly reduces the adult-child gap relative to the dense baseline, matching Experiment 1: both MoE variants give lower adult WER (by 0.10-0.17 pp) and lower child WER (by 1.50-1.53 pp) than ML-Dense<sub>child</sub>. Dutch shows the same direction, lower adult WER (by 0.78-0.79 pp), lower child WER (by 1.64-2.06 pp), and a smaller gap for both MoE variants, but the gap reduction is not statistically significant. German is the clearest counter-example: both MoE variants give lower adult WER than dense, but child WER is 10.7-11.6 pp higher, so the gap significantly increases. In Polish, MoE also gives a lower adult WER, but the gap does not improve: ML-MoE<sub>LB-off</sub> is numerically worse but not significant, while ML-MoE<sub>LB-on</sub> significantly increases the gap.

**Table 4.2:** WER results on multilingual test sets, grouped by language. Subscripts denote training stage and configuration: CV = fine-tuned on Common Voice adult speech only (adult-only stage, the multilingual analogue of Stage 1 in Table 4.1); child = additionally fine-tuned on child speech from all four languages (the analogue of Stage 2); LB-on/LB-off = whether the load-balancing auxiliary loss remained active or was disabled during the MoE model’s child-speech fine-tuning. Adult and child WERs are reported in percent; Gap = Child WER – Adult WER in percentage points. Best value within each language block and column is shown in bold.

| Lang.                          | Model                     | Adult WER (%) | Child WER (%) | Gap (pp)             |
|--------------------------------|---------------------------|---------------|---------------|----------------------|
| <b>English (EN)</b>            |                           |               |               |                      |
|                                | ML-Dense <sub>CV</sub>    | <b>3.14</b>   | 25.05         | +21.91               |
|                                | ML-MoE <sub>CV</sub>      | 3.49          | 27.41         | +23.92 <sup>†</sup>  |
|                                | ML-Dense <sub>child</sub> | 3.54          | 15.17         | +11.63               |
|                                | ML-MoE <sub>LB-off</sub>  | 3.37          | 13.67         | +10.30*              |
|                                | ML-MoE <sub>LB-on</sub>   | 3.44          | <b>13.64</b>  | +10.20*              |
| <b>Dutch (NL)</b>              |                           |               |               |                      |
|                                | ML-Dense <sub>CV</sub>    | <b>2.70</b>   | 78.78         | +76.08               |
|                                | ML-MoE <sub>CV</sub>      | 2.77          | 87.22         | +84.45 <sup>†</sup>  |
|                                | ML-Dense <sub>child</sub> | 3.50          | 22.48         | +18.98               |
|                                | ML-MoE <sub>LB-off</sub>  | 2.71          | <b>20.42</b>  | +17.71               |
|                                | ML-MoE <sub>LB-on</sub>   | 2.72          | 20.84         | +18.12               |
| <b>German (DE)<sup>a</sup></b> |                           |               |               |                      |
|                                | ML-Dense <sub>CV</sub>    | 6.94          | 68.19         | +61.25               |
|                                | ML-MoE <sub>CV</sub>      | <b>6.40</b>   | 70.07         | +63.67 <sup>†</sup>  |
|                                | ML-Dense <sub>child</sub> | 8.33          | <b>38.42</b>  | +30.09               |
|                                | ML-MoE <sub>LB-off</sub>  | 7.30          | 49.13         | +41.83 <sup>†</sup>  |
|                                | ML-MoE <sub>LB-on</sub>   | 7.41          | 49.98         | +42.57 <sup>†</sup>  |
| <b>Polish (PL)<sup>b</sup></b> |                           |               |               |                      |
|                                | ML-Dense <sub>CV</sub>    | 8.98          | 125.49        | +116.51              |
|                                | ML-MoE <sub>CV</sub>      | 8.57          | 141.30        | +132.73 <sup>†</sup> |
|                                | ML-Dense <sub>child</sub> | 9.96          | <b>93.87</b>  | +83.91               |
|                                | ML-MoE <sub>LB-off</sub>  | <b>8.48</b>   | 97.23         | +88.75               |
|                                | ML-MoE <sub>LB-on</sub>   | 8.55          | 100.40        | +91.85 <sup>†</sup>  |

\* indicates that MoE significantly reduces the adult-child gap relative to the dense model at the same training stage; <sup>†</sup> indicates that MoE significantly increases the gap. Significance testing uses the gap-difference statistic  $\Delta\text{Gap}$ .

<sup>a</sup> KidsTALC validation split is used as the German child test set because the test split has no transcriptions.

<sup>b</sup> PAVSig contains children with sigmatism and isolated-word recordings; Polish child results are therefore not directly comparable to the other child corpora.

Adult WER shows a related but separate pattern. Stage 2 fine-tuning worsens adult WER for the dense model in all four languages (by up to 1.39 pp), whereas the MoE variants are flat or improve slightly in English, Dutch, and Polish, and worsen only by 0.90-1.01 pp in German. As a result, MoE has lower adult WER than dense in every language at Stage 2, whereas at Stage 1 this was only true for German and Polish; Section 5.1.2 discusses a possible reason behind this pattern.

Polish requires separate error analysis because child WER remains above 90% for all three child-adapted models, even after Stage 2 fine-tuning cuts it by 31.6 pp for the dense model and by 40.9-44.1 pp for the two MoE variants. Polish child WER can exceed 100% because WER counts insertions as errors (Equation 3.1): PAVSig references are single isolated words, so one inserted word already doubles an utterance’s error count. The errors are driven predominantly by substitutions, which account for 87-98% of errors across all five models, with deletions near zero. In ML-Dense<sub>CV</sub> and ML-MoE<sub>CV</sub>, Polish words are frequently substituted with English or German tokens (e.g. *jeże*→yes, *czapka*→chapter), indicating language confusion. After Stage 2 fine-tuning, predictions remain in Polish but contain phonetically plausible incorrect forms instead (e.g. *żaba*→szafa, *zegar*→rzeka), showing that the model no longer switches language but still confuses acoustically similar phonemes: Polish sibilant and affricate contrasts (*ż/sz*, *cz/k*, *c/dz*) are acoustically close and are often neutralised in child speech [15], and PAVSig additionally contains children with sigmatism, which further affects these contrasts.

Load balancing has little effect in the multilingual setting: the protective hypothesis from Experiment 1, that disabling load balancing lets the router specialise more freely for child speech while keeping it enabled protects adult WER, is not supported here either. Across the four languages, the only significant difference between ML-MoE<sub>LB-off</sub> and ML-MoE<sub>LB-on</sub> is in German adult WER, where LB-off is significantly lower than LB-on, the opposite of the expected protective effect; no significant difference appears for child WER or the gap in any language. As in Experiment 1, the changes are therefore driven mainly by the child fine-tuning data itself, not by the load-balancing loss.

Experiment 2, therefore, answers RQ2 with a language-dependent result rather than a consistent multilingual advantage: MoE significantly reduces the adult-child gap only for English. Dutch is directionally positive but not significant; German is a clear counter-example despite its retained adult-WER advantage; and Polish shows no gap reduction, with ML-MoE<sub>LB-off</sub> numerically worse but not significant and ML-MoE<sub>LB-on</sub> significantly worse (Section 5.2.3). The gap reduction found for English in Experiment 1, therefore, does not transfer to the other three languages.

### 4.3. Experiment 3: Expert Routing Analysis (RQ3)

This section examines RQ3. The hypothesis, motivated by the Stage 1 result in Experiment 1, is that adult and child speech activate different experts even though the router never receives an age label, and that the experts involved compute genuinely different representations. Section 4.3.1 tests this in the English-only models; Section 4.3.2 investigates whether it survives in the multilingual models, where the router must also handle four languages.

#### 4.3.1. English Routing Analysis

Three English models are compared: **EN-MoE<sub>LS</sub>**, fine-tuned on adult speech only, and **EN-MoE<sub>LB-off</sub>** and **EN-MoE<sub>LB-on</sub>**, both fine-tuned on child and adult speech. The comparison between **EN-MoE<sub>LS</sub>** and the child-adapted models tests whether exposure to child speech changes the router. The comparison between the two child-adapted models tests whether any specialisation is robust to the load-balancing setting.

##### Age Routing Specialisation

**Routing divergence.** If the router separates adult from child speech, their routing distributions should differ. Figure 4.1 tests this with the per-layer JSD between adult and child routing distributions. In **EN-MoE<sub>LS</sub>**, JSD remains close to zero at every layer. In both child-adapted models, divergence is small in absolute terms but reliably above zero and above the baseline across most layers, rising at layers 4-5 and again at layers 13-14, with the maximum at layer 14 (approximately 0.009-0.010). A slight elevation at layer 0 shows that some age-related separation appears already at the first MoE layer. The two-peak pattern appears in both **EN-MoE<sub>LB-off</sub>** and **EN-MoE<sub>LB-on</sub>** despite their different load-balancing settings,

and the bootstrap confidence intervals at the peak layers remain above zero, so the divergence is robust to the load-balancing setting and to test-set resampling.

Clearer divergence at the JSD level appears after child fine-tuning; as the expert preference delta below shows, the adult-only baseline carries a weak age-related tendency, but it is too small to move the aggregate JSD away from zero. Age-sensitive routing is therefore not inherited from pre-training but is built up during adaptation to child speech, and it concentrates at specific layers rather than spreading evenly through the encoder.

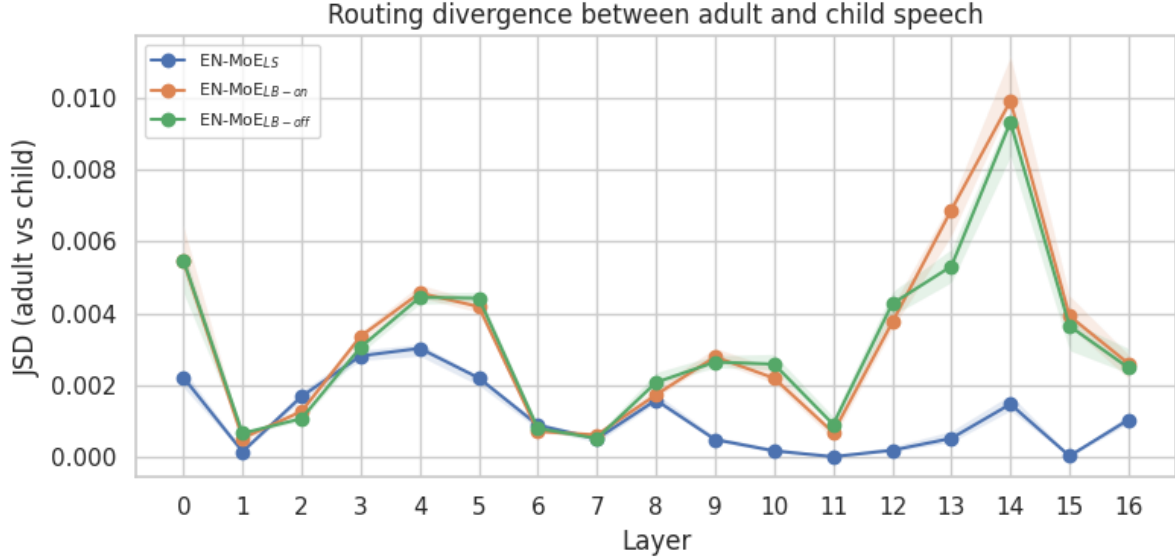


**Figure 4.1:** Jensen-Shannon divergence between adult and child routing distributions across 17 encoder layers for **EN-MoE<sub>LS</sub>** (adult-only), **EN-MoE<sub>LB-off</sub>**, and **EN-MoE<sub>LB-on</sub>**. Both child-adapted models show two divergence peaks at layers 4-5 and layers 13-14. The adult-only baseline remains near zero throughout.

**Expert preference delta.** JSD shows that routing differs, but not which experts drive the difference. The expert preference delta (child routing load minus adult routing load per expert; Equation 3.7) identifies them: positive values (red in Figure 4.2) mark child-preferred experts, negative values (blue in Figure 4.2) adult-preferred experts. In the adult-only baseline, deltas are small: E1 is mildly child-preferred and E2 mildly adult-preferred, a pattern that is visible at the expert level but modest enough that it does not move the aggregate JSD away from zero (Figure 4.1).

Child fine-tuning amplifies this pattern. In both child-adapted models, expert E2 becomes adult-preferred, strongest around layers 4-5, while expert E1 becomes child-preferred in the deeper layers, strongest around layer 14. These expert-level peaks coincide with the JSD peaks, so the divergence is driven by two experts rather than by diffuse changes across all four. Child fine-tuning, therefore, does not create an entirely new routing pattern. Instead, it strengthens and stabilises a weak tendency already present after adult-only training. This matches the Stage 1 result in Table 4.1, where the MoE encoder generalised better to child speech than the dense baseline before seeing any child data: sparse routing appears to capture some age-related acoustic variation as a property of the architecture, which child data then strengthens.

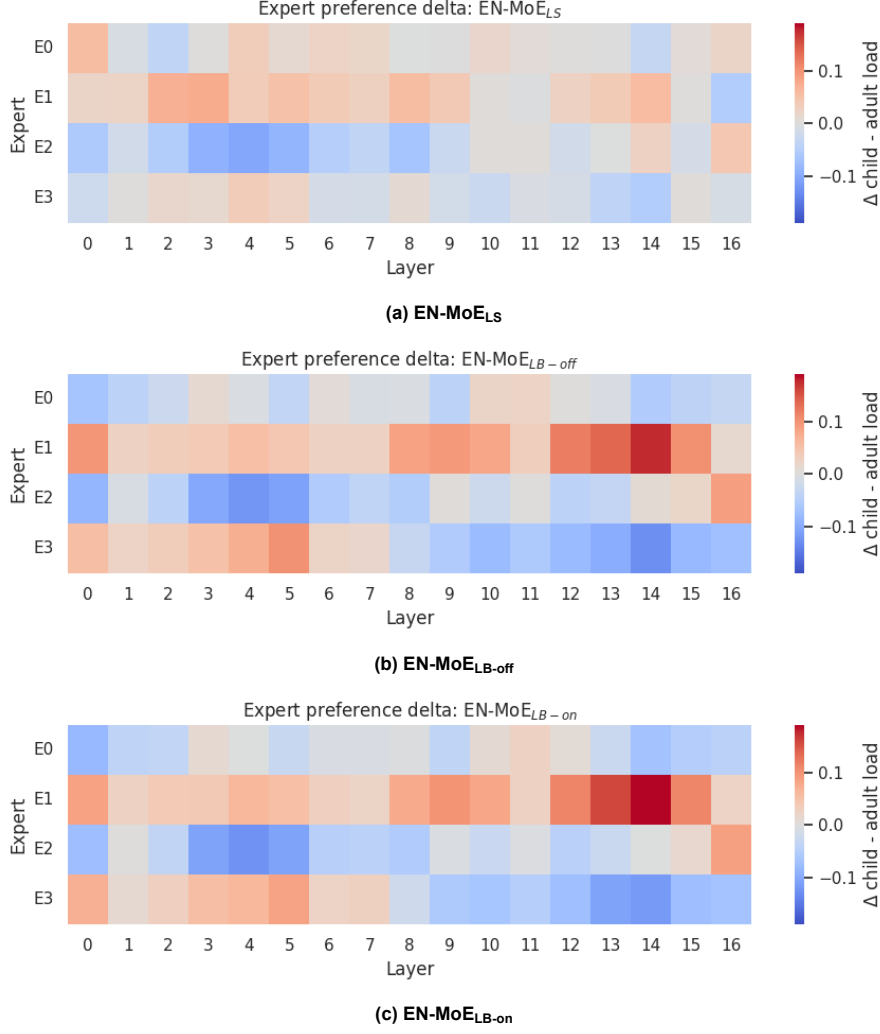


**Figure 4.2:** Expert preference delta ( $\Delta_e$ ; child routing load minus adult routing load per expert) across 17 encoder layers for the three English MoE models. Positive values (red) indicate child-preferred experts; negative values (blue) indicate adult-preferred experts. After child fine-tuning, E2 is consistently adult-preferred, and E1 becomes child-preferred in the late layers; the adult-only baseline shows the same tendencies only weakly and inconsistently.

### Functional Expert Specialisation

The router and the experts are separate components, so routing adult and child speech to different experts only matters if those experts compute different functions. If all experts produced nearly identical outputs, the routing divergence above would have no effect on the encoder representation. The expectation, if the specialisation is genuine, is that expert outputs become less similar after child fine-tuning, especially at the age-sensitive layers. To test this, pairwise cosine similarity is measured between expert outputs on the same LibriSpeech frames (Section 3.10.3) at layers 4, 5, 13, and 14, the JSD peak layers.

The results confirm this expectation (Figure 4.3). In **EN-MoE<sub>LS</sub>**, off-diagonal similarity is high, approximately 0.65-0.69 at layers 4-5 and 0.53-0.63 at layers 13-14, so the experts compute broadly comparable outputs. After child fine-tuning, similarity falls in both child-adapted models to approximately 0.45-0.59 at layers 4-5 and 0.39-0.49 at layers 13-14, with the strongest separation at layer 14, the layer with the largest routing divergence. Routing specialisation and functional specialisation therefore co-occur at the same layers: the age-related routing differences correspond to genuinely different expert computations, not to a relabelling of interchangeable experts.

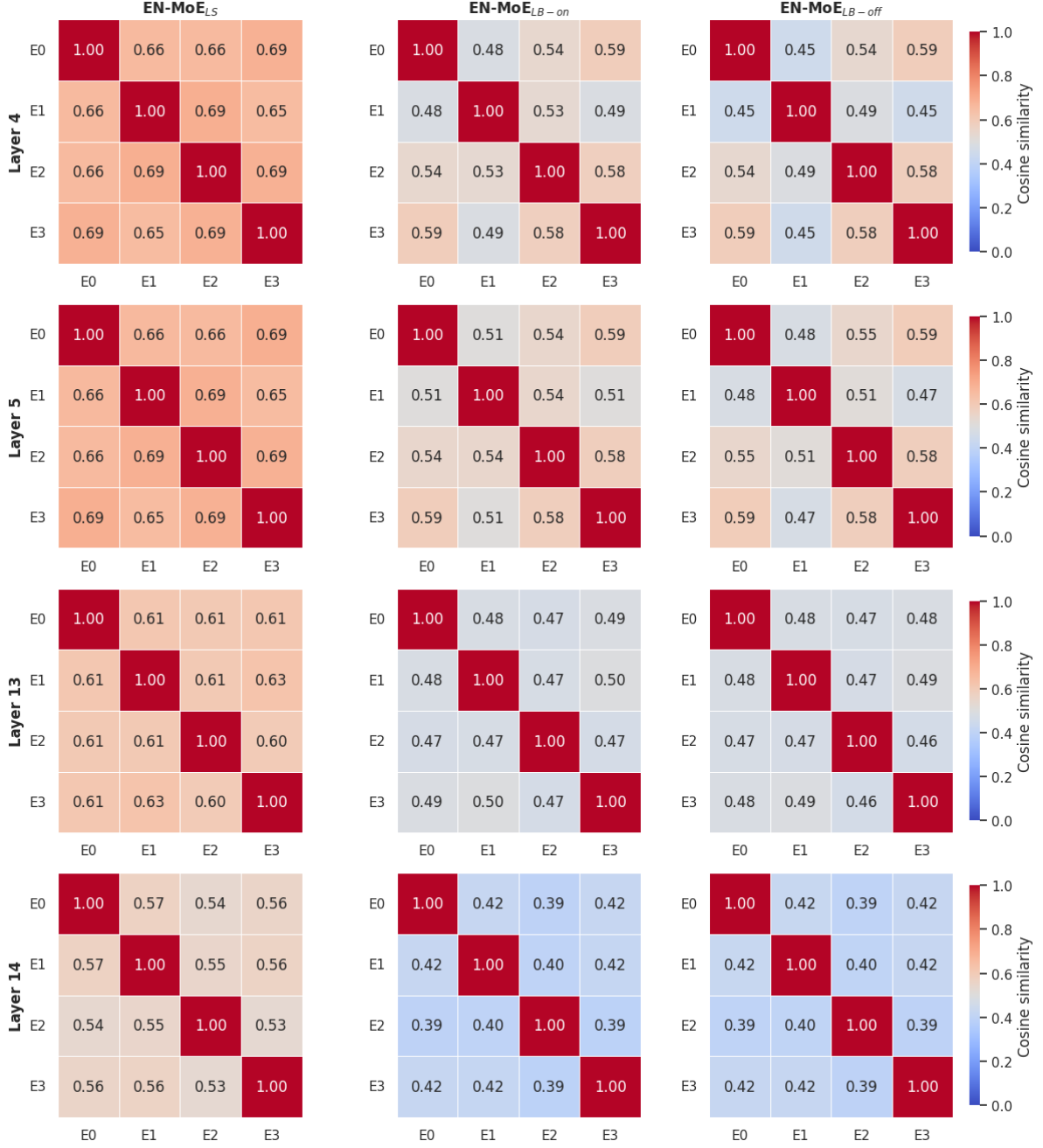


**Figure 4.3:** Pairwise cosine similarity between expert outputs at layers 4, 5, 13, and 14 (rows) for the three English MoE models (columns). Each cell shows the cosine similarity between two experts (E0-E3); diagonal cells are always 1.0 by definition. Red indicates high similarity; blue indicates that two experts produce more distinct outputs. Child-adapted models (centre and right columns) show more blue off-diagonal cells than the adult-only baseline (left), with the strongest effect at layer 14, indicating that child fine-tuning causes experts to become functionally more distinct.

### Domain Control

LibriSpeech and MyST differ not only in speaker age but also in recording conditions and speaking style, so the routing divergence above could, in principle, reflect corpus differences rather than age. Because JSD measures the divergence between two routing distributions, the control analysis is performed pairwise. Three corpus pairs are compared: LibriSpeech-MyST, where both speaker age and corpus domain differ; LibriSpeech-YODAS2, where speaker age is held constant as adult speech while corpus domain differs; and MyST-YODAS2, where both age and domain differ. If routing responded mainly to recording domain rather than speaker age, the adult-adult LibriSpeech-YODAS2 comparison should show divergence comparable to the LibriSpeech-MyST comparison.

Figure 4.4 shows the opposite. In both child-adapted models, LibriSpeech-MyST peaks at layers 4-5 and 13-14, matching the age-sensitive layers identified above, while LibriSpeech-YODAS2 remains low across the network. In the adult-only baseline, all three corpus pairs remain close to zero. Since the same corpus pairs are used for every model, divergence caused purely by corpus identity or recording conditions should also appear in the adult-only baseline. The peaks that emerge only after child fine-tuning, therefore, cannot be explained by the datasets alone.

YODAS2 is not a perfect control: it is acoustically heterogeneous, and its speaker-age composition cannot be guaranteed as cleanly as LibriSpeech or MyST. For this reason, the MyST-YODAS2 comparison is harder to interpret. The most informative comparison is LibriSpeech-MyST against LibriSpeech-YODAS2, where LibriSpeech is held constant. At the age-sensitive layers, the adult-child divergence consistently exceeds the adult-adult domain-control divergence. This supports the interpretation that the routing divergence follows speaker age more strongly than recording domain, which is necessary for the age-specialisation claim.

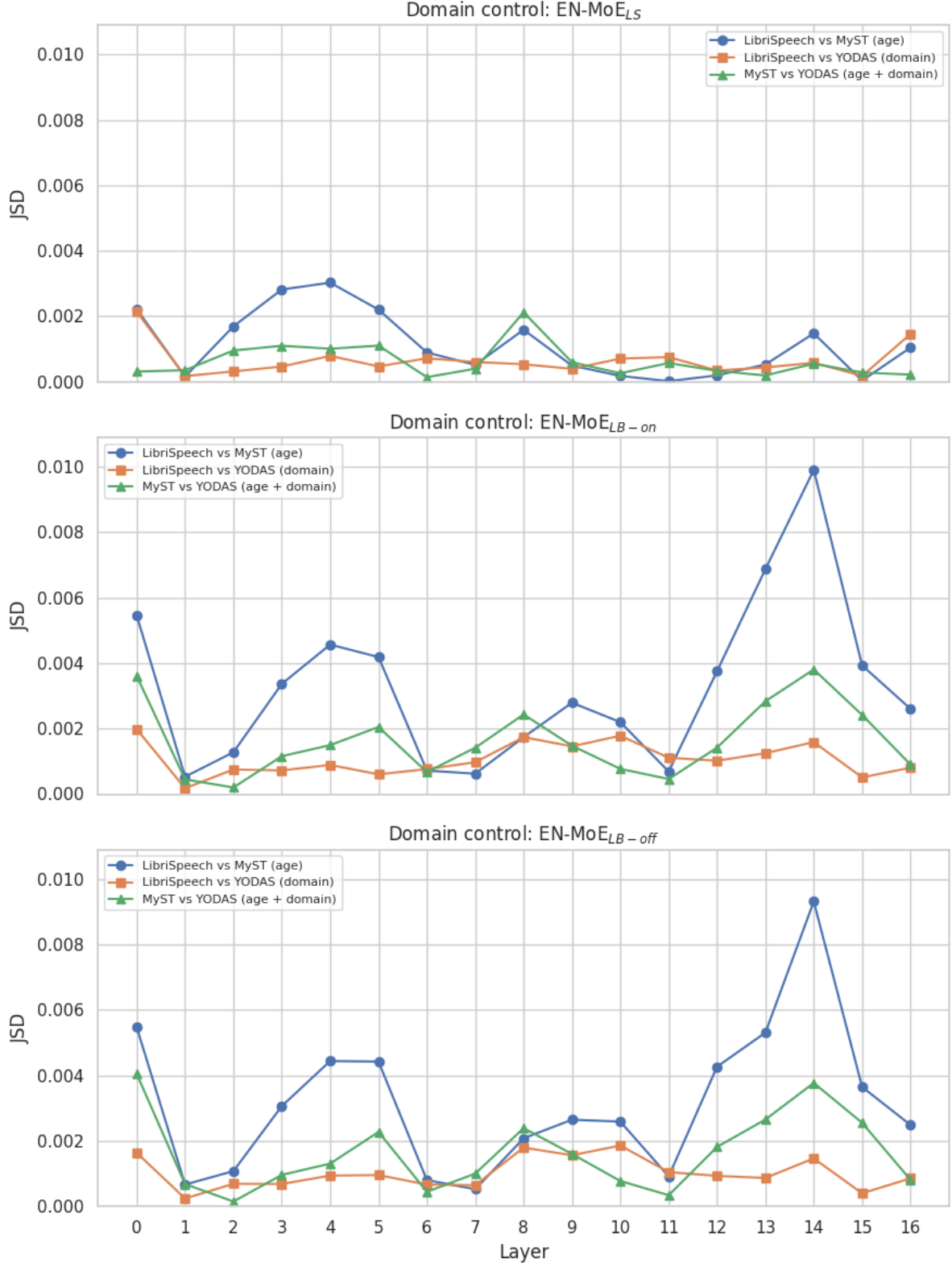


**Figure 4.4:** Domain control using YODAS2. Each subplot corresponds to one model; each line shows the JSD between the routing distributions of two corpora across 17 encoder layers. LibriSpeech-MyST (blue) contrasts adult and child speech; LibriSpeech-YODAS2 (orange) contrasts two adult corpora from different recording domains; MyST-YODAS2 (green) contrasts child and adult speech from different domains. The y-axis is shared across subplots. In the child-adapted models, the blue line peaks at layers 4-5 and 13-14 while the orange line remains low, indicating that the routing divergence at these layers is more consistent with speaker age than with the tested adult-domain contrast.

### Summary: English Routing

The English analysis confirms the RQ3 hypothesis for English. A weak age-related routing tendency exists after adult-only training; child fine-tuning amplifies it into consistent divergence at layers 4-5 and 13-14, carried mainly by an adult-preferred E2 and a child-preferred E1. At the same layers, the experts also become functionally more distinct, and the domain control shows that the divergence is more consistent with speaker age than with the tested adult-domain contrast. The router, therefore, learns to treat adult and child speech differently, with real consequences for the encoder computation, without ever receiving an age label.

### 4.3.2. Multilingual Routing Analysis

The multilingual analysis compares **ML-MoE<sub>CV</sub>**, fine-tuned on adult speech only, with the child-adapted **ML-MoE<sub>LB-off</sub>** and **ML-MoE<sub>LB-on</sub>**. This section addresses two questions in sequence. First, does the router specialise by language, the factor that prior multilingual MoE work suggests should dominate? Second, once language is held constant, does the age-related routing found for English also appear within each language?

#### Language Specialisation

**Routing divergence.** The first question is whether the router separates speech primarily by language. To isolate language from age, this analysis uses adult speech only: LibriSpeech for English and Common Voice for Dutch, German, and Polish. JSD measures divergence between two routing distributions, so it is computed pairwise between all language pairs rather than for each language in isolation. Based on prior multilingual MoE findings, clear language-dependent routing is expected across the encoder layers.

This expectation is confirmed by Figure 4.5, which plots pairwise JSD across the 17 encoder layers for each of the three multilingual MoE models, with one subplot per model and one coloured line per language pair. Cross-language JSD is close to zero in the earliest layers, rises sharply from approximately layer 5 onward, and is largest in the mid-to-late encoder, especially around layers 13-14; language-specific routing therefore emerges after the lower encoder layers rather than uniformly across the network. The strongest separation is observed for pairs involving Polish: NL-PL (purple) and DE-PL (brown), reaching approximately 0.20-0.29 in the mid-to-late layers. By contrast, EN-DE (orange) and EN-NL (blue) remain among the least divergent pairs throughout. This ordering is consistent with the Germanic languages, English, Dutch, and German, being phonologically closer to each other than to the Slavic Polish, although the analysis does not test which linguistic properties drive the routing differences. One caveat is that the English adult data comes from LibriSpeech (read audiobooks) while the Dutch, German, and Polish adult data come from Common Voice, so pairs involving English mix language differences with corpus-source differences; the comparisons among the three Common Voice languages, including the strong Polish separation, are not affected by this confound. The overall profile is consistent across all three multilingual MoE models.

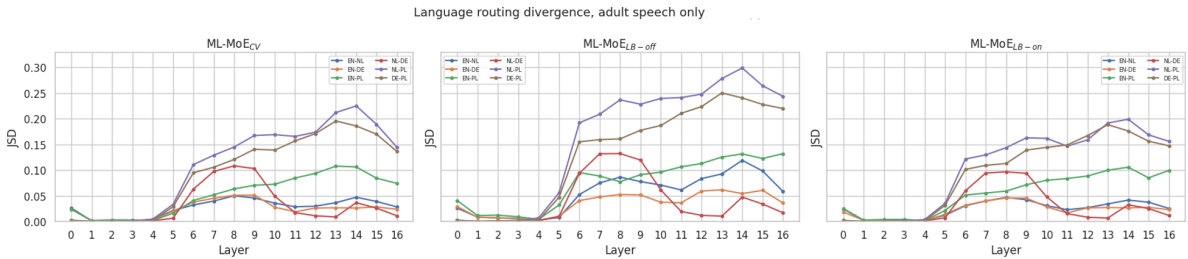


**Figure 4.5:** Cross-language routing divergence (JSD) for adult speech across 17 encoder layers in the three multilingual MoE models (4-hour capped samples). JSD quantifies the divergence between the routing distributions of two corpora and cannot be computed for a single language in isolation, so results are shown as pairwise comparisons across all six language pairs. Each subplot corresponds to one model; each line corresponds to one language pair. The y-axis is shared across subplots.

**Language-related expert preference delta.** The expert-load heatmaps in Figure 4.6 show which experts carry the language separation observed in the pairwise JSD analysis. For each language,

expert, and layer, the plotted value is the deviation from the uniform top-2 routing baseline. Because the multilingual model has 8 experts and each token is routed to 2 experts, uniform routing would assign each expert a load of  $2/8 = 0.25$ . Red cells therefore indicate experts that receive more tokens than expected under uniform routing, while blue cells indicate experts that receive fewer tokens. Values close to zero indicate no strong language-specific preference for that expert at that layer.

Language specialisation is expert-specific rather than distributed evenly across all experts. Polish shows the strongest and most stable pattern: from approximately layer 5 onward, it routes substantially more than the uniform share to expert E2, while experts E0 and E3 are underused. Dutch instead favours expert E1 across much of the network and underuses expert E2 in the mid-to-late layers. German shows a more moderate preference for experts E0 and E1 in parts of the encoder, while English remains comparatively diffuse, with no single expert dominating across layers. These expert-load patterns explain why language pairs involving Polish have the highest JSD: Polish is not only different in its aggregate routing distribution, but is consistently concentrated on a different expert.

These language-expert associations remain visible after child fine-tuning, especially the Polish preference for expert E2 and the Dutch preference for expert E1. Child fine-tuning, therefore, does not remove the language-specialised routing structure learned during adult multilingual training. Language remains a strong routing signal in the multilingual encoder, which motivates the next analysis: whether an age-related signal can still be detected within each language.

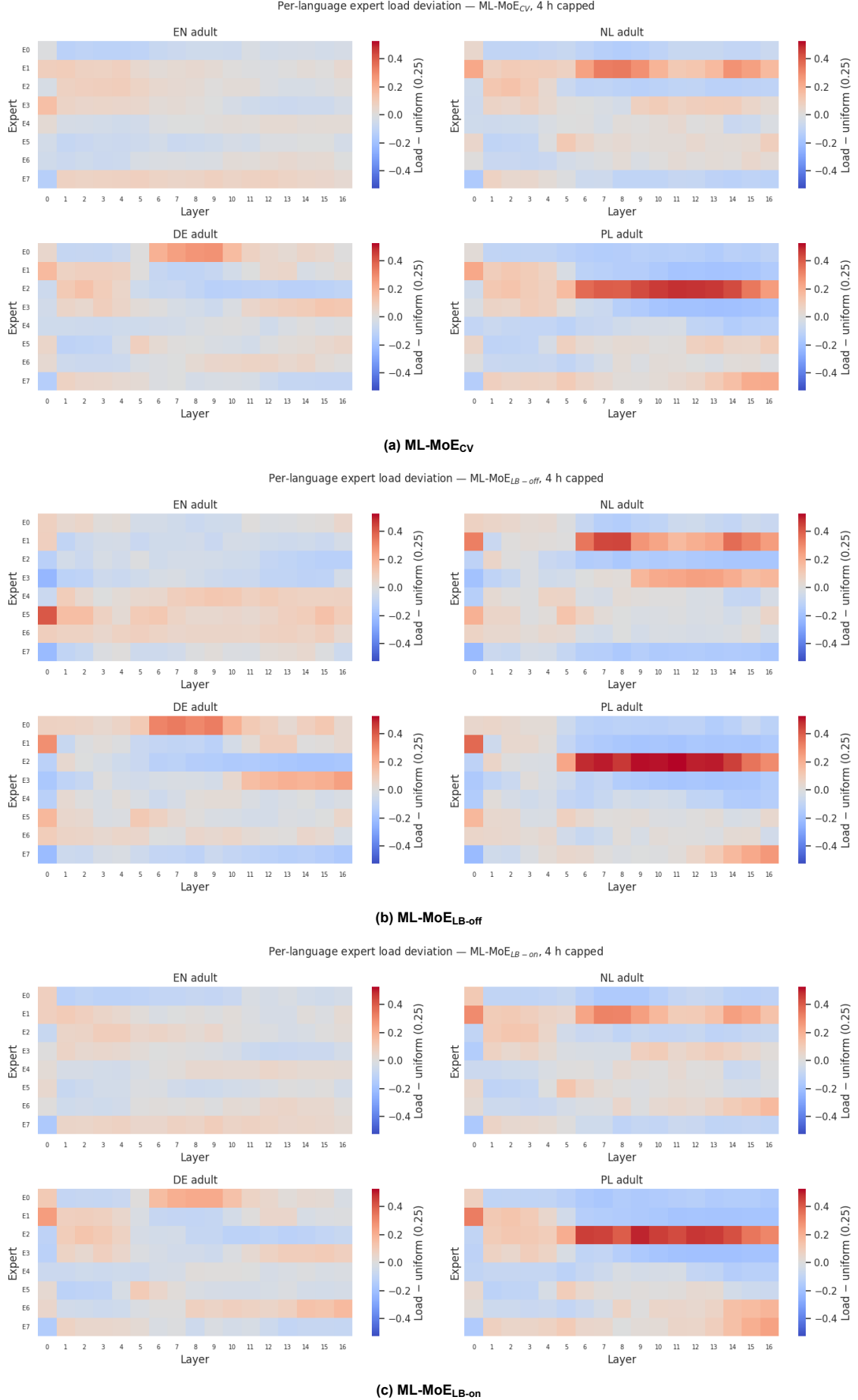


**Figure 4.6:** Per-language expert load deviation from the uniform baseline (0.25) for adult speech in English (EN), Dutch (NL), German (DE), and Polish (PL). Red cells indicate that the expert receives more tokens than the uniform share of 0.25; blue cells indicate it receives fewer. Polish shows sustained above-uniform load on E2 from mid-layers onward; Dutch preferentially uses E1; German moderately activates E0; English is diffuse. The language-expert associations are consistent across all three models.

### Age Specialisation Within Language

**Routing divergence.** If the English age finding generalises, adult and child speech should route differently within each language of the multilingual model, although more weakly than the language signal. This analysis is carried out for English, Dutch, and German. Polish is reported separately in this section, since PAVSig shows very high WER and language-confusion errors (Section 4.2) and is not a typical Polish child-speech condition.

Figure 4.7 shows within-language adult-child routing divergence for English, Dutch, and German. The columns correspond to languages, and the rows correspond to models: the adult-only model **ML-MoE<sub>CV</sub>** is shown in the top row in purple, while the two child-adapted models, **ML-MoE<sub>LB-off</sub>** and **ML-MoE<sub>LB-on</sub>**, are shown in the middle and bottom rows in blue and green. Each subplot shows JSD across the 17 encoder layers for one language-model combination.

The age signal is present but clearly weaker than the language signal. Within-language adult-child JSD remains below approximately 0.06, much lower than the cross-language JSD of approximately 0.20-0.29 reported above. The clearest age-related routing appears in Dutch and German. Dutch shows elevated divergence in the late layers, especially for the adult-only model and **ML-MoE<sub>LB-off</sub>**, while German shows a broader mid-to-late-layer increase after child fine-tuning. English remains comparatively weak, with only small increases after child adaptation. Unlike the English-only model, where age-related divergence forms two clear peaks at layers 4-5 and 13-14, the multilingual profiles are less localised and often begin with elevated divergence at layer 0. Age-related routing is therefore detectable in the multilingual model, but it is weaker than language routing and varies substantially by language.

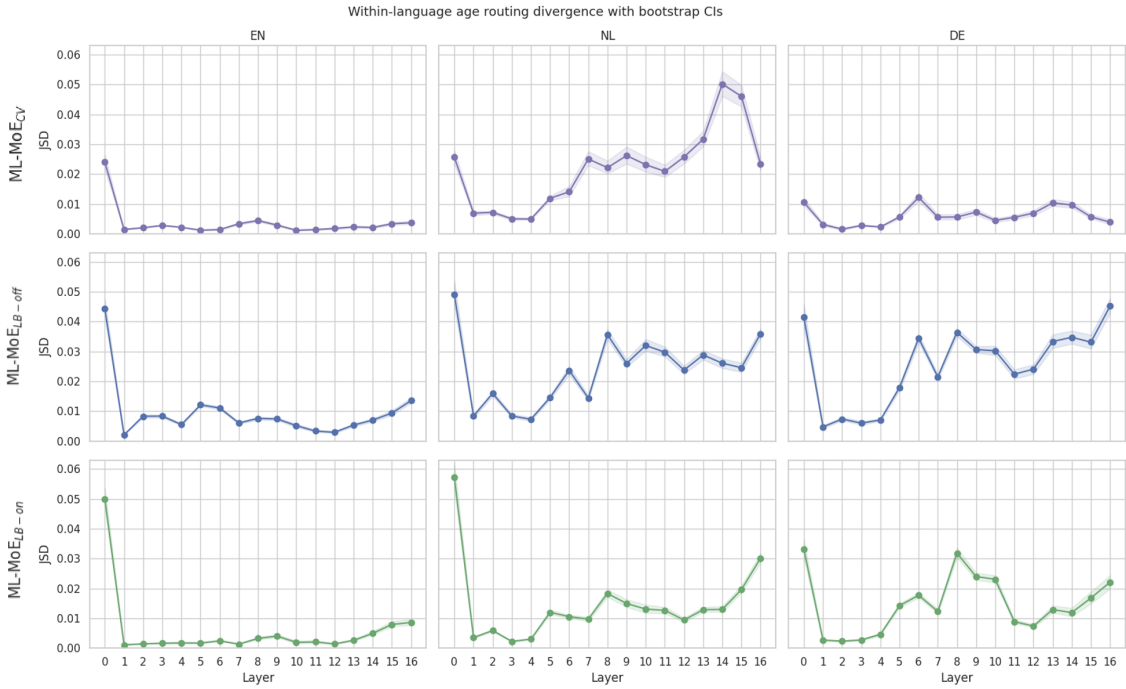


**Figure 4.7:** Within-language adult-child routing divergence (JSD) across 17 encoder layers. Rows are models; columns are languages (EN, NL, DE). Shaded regions show 95% bootstrap confidence intervals. To compare, read across rows within a column: the child-adapted models (rows 2-3) show higher JSD than the adult-only baseline (row 1) across mid-to-late layers, most clearly in the DE column and across a broader range of layers in the NL column.

**Age-related expert preference delta.** Figure 4.8 shows which experts carry the within-language age differences. In Figure 4.8 rows correspond to models and columns correspond to languages. Each cell shows the expert preference delta, defined as child routing load minus adult routing load for one expert at one layer (equation 3.7). Red values indicate child-preferred experts, blue values indicate adult-preferred experts, and values close to white indicate little difference between child and adult routing.

The patterns are partly shared across languages and partly language-specific. In the child-adapted

models, expert E3 becomes child-preferred in several languages, most clearly for German in **ML-MoE<sub>LB-off</sub>**, where the delta exceeds 0.3 near the final layers. The same expert shows a weaker child preference in English and Dutch. Adult-preferred routing is more visible for expert E4, expert E5, and expert E6, especially in Dutch and German. The clearest language-specific pattern is Dutch, where expert E1 becomes strongly child-preferred in **ML-MoE<sub>LB-off</sub>**, a pattern not seen as clearly in English or German.

The two load-balancing variants differ in strength, but they show the same general conclusion: age-related routing is not carried by one universal child-speech expert across all languages. Instead, child and adult preferences appear within the language-specialised expert structure already present in the multilingual encoder.

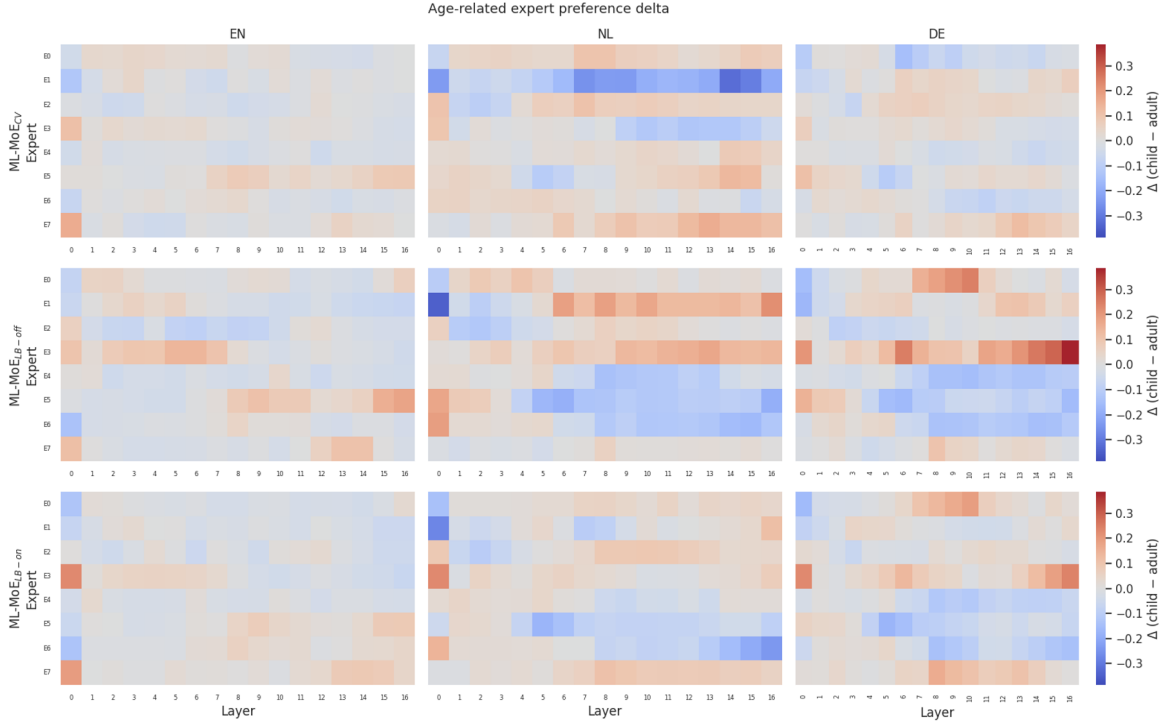


**Figure 4.8:** Age-related expert preference delta (child routing load minus adult routing load per expert; Eq. 3.7) for the multilingual MoE models. Rows are models; columns are languages (EN, NL, DE). Red = child-preferred; blue = adult-preferred. Child-adapted models show stronger and more structured preferences than the adult-only baseline.

#### Polish (PAVSig) Analysis

**Routing divergence.** Figure 4.9 shows the routing divergence between adult Polish Common Voice speech and PAVSig across the 17 encoder layers. The purple line corresponds to the adult-only model, **ML-MoE<sub>CV</sub>**, while the blue and green lines correspond to the child-adapted models, **ML-MoE<sub>LB-off</sub>** and **ML-MoE<sub>LB-on</sub>**. Unlike the English, Dutch, and German analyses above, this comparison is not interpreted as typical age specialisation. Instead, it shows how differently PAVSig is routed from adult Polish speech.

The adult-only model already shows high divergence, peaking at approximately 0.11 around layer 10. Since this model has not seen child speech, the divergence cannot reflect learned child-speech specialisation. Rather, it suggests that PAVSig is already far from the adult Polish routing distribution before child adaptation. This is consistent with the error analysis in Section 4.2, where the adult-only models often substitute Polish words with English or German tokens.

After child fine-tuning, the divergence drops to approximately 0.01-0.05 for most layers. This is the opposite direction from the English, Dutch, and German analyses, where child adaptation generally increases adult-child routing divergence. For PAVSig, child fine-tuning therefore appears to reduce the

mismatch with adult Polish routing. This matches the post-adaptation error pattern: the outputs remain Polish rather than switching to English or German, although WER remains very high.

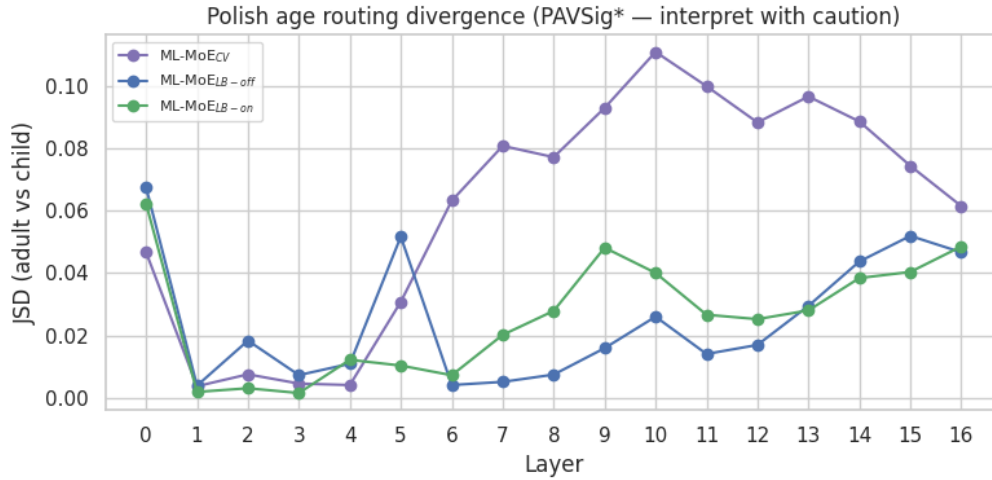


**Figure 4.9:** Polish adult-child routing divergence using PAVSig as the child speech set. JSD is computed between the adult Polish Common Voice speech and the Polish PAVSig speech across the 17 encoder layers for the three multilingual MoE models. PAVSig contains isolated-word utterances from children with sigmatism, so this comparison does not represent typical Polish child speech. The high divergence in **ML-MoECV** indicates that PAVSig is already far from the adult Polish routing distribution before child-speech adaptation.

**Expert preference delta.** Figure 4.10 shows which experts drive the difference between PAVSig and adult Polish speech. The plotted value is the expert preference delta, computed as PAVSig routing load minus adult Polish routing load. Red values indicate PAVSig-preferred experts, blue values indicate adult-Polish-preferred experts, and values close to white indicate little difference between the two routing distributions.

The adult-only model shows a strong negative delta for expert E2 across the middle and later layers. This means that adult Polish speech uses expert E2 much more than PAVSig does. This is important because expert E2 is also the main Polish-associated expert in the language-specialisation analysis (Figure 4.6). The PAVSig mismatch, therefore, appears both in the high JSD values and in the reduced use of the main adult-Polish expert.

After child fine-tuning, the expert E2 contrast becomes weaker and less consistent. This supports the same interpretation as the JSD result: child adaptation makes PAVSig routing more similar to adult Polish routing, but it does not produce a clear typical Polish child-speech routing pattern. The Polish results are therefore best read as a PAVSig mismatch analysis rather than as evidence for or against age specialisation in typical Polish child speech.

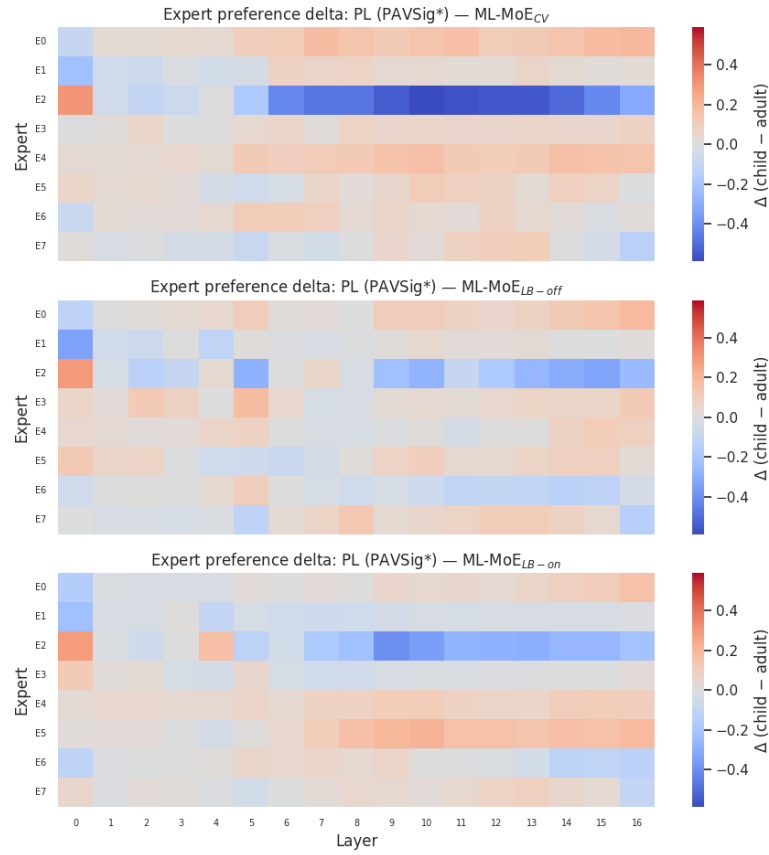


**Figure 4.10:** Expert preference delta for Polish PAVSig vs adult Polish speech across the three multilingual MoE models. Red = PAVSig-preferred; blue = adult Polish-preferred. The adult-only model shows a large routing shift (strong negative delta for E2); the child-adapted models show weaker, less consistent patterns. Because PAVSig is isolated pathological speech, these deltas are not interpreted as typical Polish child-speech preferences.

**Summary: Multilingual Routing**

The multilingual routing analysis supports RQ3 more cautiously than the English-only analysis. Language is the dominant routing signal: adult speech from different languages activates clearly different expert patterns, and this language-specialised structure remains visible after child fine-tuning. A weaker age-related signal is layered on top of it, without requiring an explicit age label.

Within language, age-related routing differences are still detectable for English, Dutch, and German, most clearly for Dutch and German. These differences do not point to one universal child-speech expert; instead, child- and adult-preferred experts depend on the language. Polish is treated separately, since PAVSig is not a typical Polish child-speech condition and its routing pattern reflects a PAVSig-adult Polish mismatch rather than age specialisation.

This evidence is weaker than the English result: it is entangled with language routing rather than isolated from it, and it is not backed by a functional expert analysis, which was carried out only for the English models.

# 5

## Discussion

This thesis investigated whether sparse Mixture-of-Experts (MoE) layers can reduce the adult-child WER gap in automatic speech recognition, and whether any reduction generalises across languages. The results in Chapter 4 give a positive answer to RQ1: the MoE encoder reduces the gap for English by a modest but statistically significant margin, and part of this advantage is present before any child speech audio is used for fine-tuning. The answer to RQ2 is language-dependent: in the multilingual setting, the gap reduction is significant only for English, Dutch shows the same direction without significance, German is a counter-example where the gap widens, and the Polish results cannot be read as typical child speech (Section 5.2.3). For RQ3, the routing analysis shows that age-related expert specialisation emerges without demographic labels: clearly in the English models, and more weakly, entangled with language routing, in the multilingual model. This chapter interprets these findings (Section 5.1), discusses their limitations (Section 5.2), outlines future work (Section 5.4), and states the contributions of the thesis (Section 5.5).

### 5.1. Discussion of the results

#### 5.1.1. Does the MoE encoder reduce the adult-child gap in English ASR? (RQ1)

The MoE encoder reduces the adult-child WER gap in English relative to the dense model, although the margin is modest (Table 4.1). This reduction does not come from sacrificing adult performance: both adult and child WER improve under the MoE encoder. The Stage 1 comparison carries most of the interpretive weight for RQ1, because **EN-MoE<sub>LS</sub>** and **EN-Dense<sub>LS</sub>** were both trained only on adult LibriSpeech at that point. The smaller gap produced by **EN-MoE<sub>LS</sub>**, therefore, cannot be explained by exposure to child-speech fine-tuning data, and the advantage persists, at a smaller margin, after Stage 2 child fine-tuning.

The high Stage 1 child WER of both models reflects two forms of mismatch at once: the acoustic differences between adult and child speech described in Chapter 1, and the domain mismatch between MyST, which contains spontaneous conversational speech, and LibriSpeech, which contains read speech [6], [10]. The MoE encoder handles this combined mismatch better than the dense model despite identical training data. This suggests that the advantage is at least partly architectural.

Two architectural properties could explain the improvement: input-conditional routing and the larger pool of total parameters in the MoE encoder. The WER results alone cannot separate these two properties, because both differ between the dense and MoE encoders. The two architectures are matched in active parameters, but not in total parameters (Table 3.4), so the larger stored parameter pool is the relevant alternative explanation, not a general capacity advantage. The routing analysis supports the routing explanation: a weak age-related routing tendency is already present after adult-only training, and child-speech fine-tuning amplifies it. This pattern is consistent with routing, rather than the larger total parameter pool alone, contributing to the English gap reduction.

The load-balancing auxiliary loss appears slightly harmful during child-speech fine-tuning. A plausible

explanation is that it penalises unequal expert utilisation, which may limit the router’s ability to specialise in child speech acoustics. Without this constraint, the router can specialise more freely for child speech. The Stage 2 differences point in the direction predicted by this explanation, but they are not statistically significant.

### 5.1.2. Does the MoE encoder reduce the adult-child gap across languages? (RQ2)

The gap reduction found for English does not generalise uniformly across languages: the reduction is significant only for English, directionally positive but not significant for Dutch, reversed for German, and indeterminate for Polish. The cross-language pattern does not have a single clean explanation: both the amount of child training data and the domain match between the adult and child training data are relevant. MyST is the largest child corpus in the training mix by a wide margin, and English is also the language in which the MoE encoder gains the most over the dense model. For Dutch and German, the child corpora are substantially smaller, but corpus size alone does not explain the difference between them: JASMIN is slightly smaller than KidsTALC (Table 3.2), yet the gap reduction is directionally positive for Dutch and significantly worse for German. Dutch also has the smallest adult training corpus of the four languages (Table 3.1); this additional data limitation may explain why the positive direction seen for Dutch does not reach significance, though the present experiment cannot separate corpus size from test-set variability or other language-specific effects.

The main difference for German is a domain mismatch. KidsTALC provides very little training data (Table 3.2) and consists of semi-spontaneous elicited speech, including picture description, narration, and examiner-child conversation, while Common Voice DE, the adult German training corpus, is read speech. This mismatch is large relative to the corpus size. A similar but smaller mismatch exists for Dutch: JASMIN consists of read-aloud texts and human-machine dialogues rather than spontaneous conversation, which brings its recording style closer to the read/prompted speech in Common Voice NL than KidsTALC’s fully spontaneous format is to Common Voice DE. With so little child data and a large acoustic domain gap between the child and adult corpora, the router cannot establish reliable German child-speech routing patterns from the available data, and the added parameters of the MoE encoder increase overfitting risk rather than providing useful specialisation. The dense model, which does not route by input, is less sensitive to this combined data and domain problem. Whether corpus size, domain mismatch, or their combination is the primary factor cannot be determined without additional training runs outside the scope of this thesis.

Whether the MoE encoder helps or hurts for typical Polish child speech cannot be determined from PAVSig alone, as discussed in Section 5.2.3.

Several controlled choices in the training setup make the architecture comparison fair, but they also limit how purely the multilingual results can be interpreted as architectural effects. The fixed-step design was chosen to compare the dense and MoE encoders under the same optimisation budget, with the same number of backward passes, at each training stage. This design answers whether the MoE encoder outperforms the dense model given equal training effort, but it does not test whether either architecture, or any individual language, has fully converged within that budget.

The first such choice concerns language balancing. In Stage 1, the multilingual training batches were language-balanced: English, Dutch, German, and Polish each contributed a fixed 25% of every training batch. This prevented the larger corpora, such as the English LibriSpeech, from dominating the multilingual training objective and ensured that all four languages received the same nominal share of updates. However, because both multilingual models are initialised from a checkpoint pretrained only on English speech, equal sampling does not mean that all languages started from the same point: the encoder already carried some acoustic bias toward English before multilingual fine-tuning began. Dutch, German, and Polish may therefore need more training steps than English to reach a comparable degree of convergence, which cannot be tested under the fixed-step design used here. This applies to both architectures equally, so it could contribute to the English-favouring pattern found across RQ2 independently of any architectural difference between the MoE and dense models.

A second choice concerns the number of experts per each MoE block used for the multilingual model, which affects the comparison between Experiment 1 and Experiment 2. The multilingual MoE encoder has twice as many experts per layer as the English-only MoE encoder, and each language still receives

only a fixed 25% share of every training batch. The multilingual models are trained for twice as many Stage 1 steps as the English-only models, but each language receives only a 25% share of every batch, so English is optimised on roughly half as many English batches as in Experiment 1. With eight experts to train instead of four, and less per-language data per step, the multilingual router receives less per-language training signal than its English-only counterpart despite the larger overall step budget. The adult-only multilingual comparison, therefore, cannot be read as a direct replication of the English-only Stage 1 setting.

A third choice concerns Stage 2 training length. Because both stages use a fixed number of optimisation steps, the setup gives a fair comparison between architectures under the same training budget; this is why the WER differences reported in Section 4.2 can be attributed to architecture rather than to one model simply training longer than the other. However, this same design cannot separate the effect of child-speech fine-tuning from the effect of continuing training for more steps in general. The MoE encoder’s adult WER stays flat or improves slightly from the adult-only to the child stage in three of the four languages, while the dense model’s adult WER worsens in every language. A natural reading of this pattern is that the MoE encoder protects adult performance better than the dense model when child speech is added. That reading assumes both architectures have converged by the end of Stage 1, so that the remaining difference reflects architecture rather than training budget, which is not guaranteed given the larger parameter count of the multilingual MoE encoder. If it had not fully converged after Stage 1, the additional gradient steps taken during Stage 2 could improve adult WER on their own, independently of any benefit from the child speech data itself, producing the same qualitative pattern without reflecting genuine robustness to catastrophic forgetting. Because Stage 1 and Stage 2 differ in both training duration and data composition at the same time, the two explanations cannot be separated with the data collected for this thesis.

These results therefore answer RQ2 under a fixed training budget shared across languages and architectures: this makes the dense-versus-MoE comparison fair, but it does not fully separate architecture from convergence, initialisation, and data or domain effects. Disentangling these effects from genuine architectural differences would require additional training runs. For the initialisation bias, training from a multilingual rather than an English-only checkpoint, or from random initialisation, with equal amounts of data per language, would remove the English-favouring starting point; training each language for longer than the fixed step budget used here would test whether the gap closes with more steps alone. For the router-capacity limitation, more Stage 1 steps or fewer experts would isolate the effect of the per-language training signal. For the adult-WER-protection comparison, a control that continues Stage 1 adult-only training for as many additional steps as Stage 2, without introducing child speech, would separate training duration from the effect of the child data itself. All of these controls fall outside the scope of this thesis and are left for future work.

### 5.1.3. Do experts specialise without demographic supervision? (RQ3)

The routing analysis gives a positive answer to RQ3. After child-speech fine-tuning, both the English and multilingual MoE models learn routing patterns that are related to speaker age, even though no demographic labels are used during training. The Stage 1 WER advantage of **EN-MoE<sub>LS</sub>** over **EN-Dense<sub>LS</sub>** suggests that the MoE architecture itself provides some benefit for out-of-distribution speakers; child-speech fine-tuning then amplifies this into the stable, visible routing patterns reported in the results.

The distribution of JSD peaks across encoder layers is also informative. Prior work on end-to-end ASR encoders suggests that earlier layers encode more local acoustic and phonetic information, whereas later layers increasingly capture more abstract and task-specific representations [40]. Child speech differs from adult speech in several low-level acoustic properties: children have higher and more variable fundamental frequency, higher formant frequencies due to shorter vocal tracts, and less stable articulation [6], [14], [15]. The JSD peak at layers 4-5 is therefore consistent with the router responding to acoustic differences such as pitch, spectral shape, and fine temporal structure. However, the routing analysis does not identify which specific features cause the divergence.

A second peak appears later, at layers 13-14. Later encoder layers are more likely to reflect increasingly abstract phonetic, phonological, and task-relevant information [40]. Child speech can also differ from adult speech at these levels, for example through consonant-cluster simplification and variable

fricative production [6], [14], [15]. The later peak is therefore consistent with the router responding to higher-level speech differences, although the present analysis cannot determine the exact properties involved. Depth-dependent routing has also been observed in large MoE language models such as Mixtral [71], although that work concerns text models and different forms of specialisation, so the comparison is limited to the general observation that routing behaviour can vary across depth. Because the OmniRouter uses the same routing parameters at every encoder layer, the depth-specific peaks arise from differences in the hidden representations presented to that shared router. They therefore indicate the layers at which adult and child representations lead to the strongest routing separation, rather than differences between independently trained layer-specific routers.

The expert preference analysis shows that the routing divergence is not spread evenly across all four experts. Instead, it is driven by a small number of experts who become consistently preferred. After child fine-tuning, expert E2 becomes mostly adult-preferred across many layers, while expert E1 becomes child-preferred in the later part of the network.

Routing divergence is only interpretable if the experts that receive different inputs also compute different representations. If all experts produced the same output regardless of which tokens they received, routing differences would have no effect on what the model actually encodes. The cosine similarity analysis provides this evidence: after child-speech fine-tuning, pairwise cosine similarity between expert outputs decreases strongly, with the largest decrease at layer 14, which is also the layer with the highest JSD. The experts are not only routing child and adult speech differently; the experts that receive them also transform the input differently.

LibriSpeech and MyST datasets differ not only in speaker age but also in recording conditions, speaking style, and corpus origin, so the routing differences could, in principle, reflect these dataset differences rather than age itself. To test this, three pairwise comparisons are made across LibriSpeech, MyST, and YODAS2: LibriSpeech-YODAS2 isolates domain shift between two adult corpora, while LibriSpeech-MyST and MyST-YODAS2 both involve the age shift, but against different domain backgrounds. YODAS2 consists of unscripted YouTube speech and is not pure read speech; its speaking style plausibly resembles MyST’s spontaneous, conversational recordings more closely than it resembles LibriSpeech’s read audiobooks. If domain or speaking style were the main driver of the routing divergence, YODAS2 should therefore route more like MyST than like LibriSpeech. Instead, LibriSpeech-YODAS2 divergence stays well below LibriSpeech-MyST divergence at the age-sensitive layers, even though YODAS2 is stylistically the closer match to spontaneous child speech of the two adult corpora. This strengthens the case for an age-related explanation: the routing difference tracks whether the speaker is a child, not whether the speech is scripted or spontaneous, although these three corpus pairs cannot rule out every possible corpus or elicitation difference between LibriSpeech and MyST.

In the multilingual models, routing divergence correlated with language identity is substantially larger than routing divergence correlated with speaker age. This language specialisation is also depth dependent: cross-language routing divergence is close to zero in the earliest layers, rises from approximately layer 5 onward, and is largest in the mid-to-late encoder. Language specific routing, therefore, emerges only after the lower encoder layers, rather than being distributed uniformly across the network. Language identity is a higher-level property of speech rather than a low-level acoustic one, so it is consistent with Belinkov and Glass [40] that this separation strengthens in the deeper, more abstract layers of the encoder rather than the shallow layers that encode raw acoustic detail.

Different languages preferentially activate different experts: The ordering of the language pairs also matches the language-selection rationale in Chapter 1: routing divergence is smallest among the Germanic languages, English, Dutch, and German, and largest for pairs involving Polish, whose phonology, including its rich sibilant system, differs most from the other three. This Germanic-Slavic reading is a plausible interpretation, but the routing analysis does not test which linguistic properties drive the ordering, so it is not an established cause.

This language-specific structure is largely preserved after child-speech fine-tuning. Age-related routing divergence is also detectable within individual languages, most clearly for Dutch and German; English shows a less consistent pattern. The multilingual encoder therefore does not use one universal child-speech expert; age-related routing is instead expressed through the existing language-specific structure, consistent with findings in multilingual MoE models where language-specific specialisation

emerges without explicit supervision [19], [23].

The PAVSig routing results show a distinct pattern. Before child-speech adaptation, the adult-only model already shows the highest routing divergence from typical adult Polish of any comparison in this analysis. PAVSig speech, which combines isolated-word recordings with pathological sibilant production, is acoustically far from the fluent read speech in Common Voice Polish that the model was trained on; both the recording format and the sibilant distortions are plausible contributors to this divergence. After child-speech fine-tuning, the divergence decreases: PAVSig is routed more similarly to adult Polish, consistent with the error analysis showing that child-adapted models produce phonetically plausible Polish output rather than cross-language substitutions. Child fine-tuning therefore reduces the routing mismatch for PAVSig, though whether this reflects better handling of pathological acoustics or simply increased Polish exposure cannot be determined from these results alone.

The routing analysis identifies where and how age-related specialisation emerges in the MoE encoder, but the exact acoustic features driving these routing decisions remain an open question. Which properties of child speech, such as formant frequencies, fundamental frequency, or articulation patterns, most strongly influence routing at layers 4-5 and 13-14 would require acoustic probing of individual expert outputs, and is a natural direction for follow-up work.

## 5.2. Limitations

### 5.2.1. Data limitations

The data limitations described here directly constrain how far the results of this thesis can be generalised. They explain why child WER remains high even after fine-tuning, why the MoE advantage is clearest for English and less consistent for other languages, and why the routing analysis can be carried out more thoroughly for English than for Dutch, German, or Polish.

The four child speech corpora used in this work reflect common limitations in child speech data collection. Collecting speech from children requires ethics approval, consent from parents or guardians, and cooperation from schools or clinics. This makes data collection much harder than for adult speech, which can often be collected from volunteers or taken from broadcast media.

As a result, the child datasets used here are much smaller than the adult datasets. Even MyST, the largest child corpus, is more than two orders of magnitude smaller than the English adult data used for pre-training. The non-English child corpora are smaller still, each under nine hours of speech (Table 3.2) and a small fraction of the corresponding adult training data. This low-resource setting limits how much the model can adapt to child speech without overfitting or losing adult performance. This limitation exists regardless of the model architecture.

Privacy is another important constraint. Speech from identifiable speakers, especially minors, cannot usually be shared freely. This limits the ability of the research community to build large child speech datasets over time and makes community-driven data collection approaches, common for adult speech, difficult to apply to child speech.

The design of each corpus also limits what kind of speech is represented, for both adult and child speakers. Each corpus was collected through a specific recording or elicitation format, such as audiobook narration, isolated-word recording, picture-description and narrative tasks, or topic-constrained tutoring interactions, rather than capturing the full range of ways people naturally speak. None of the corpora covers the full range of speaking styles, topics, and prosodic patterns found in natural adult or child speech. The WER results, therefore, reflect performance on these specific speech styles and may not transfer to speech that falls outside them, such as unscripted conversation, storytelling, or spontaneous interaction in everyday settings.

A further uncertainty concerns the gender composition of the adult training data. As described in Section 3.2.3, most Common Voice speakers report no gender, and among those who do, male speakers greatly outnumber female speakers. If the underlying Dutch, German, and Polish adult corpora are indeed male-skewed, they would underrepresent the higher fundamental frequencies typical of female adult speech, which lie between typical male adult speech and child speech; adult-only training would then provide the encoder with less exposure to the region of the acoustic space closest to child speech.

Because the true gender composition is unknown, this effect cannot be quantified, and it is noted as an unresolved property of the training data rather than a measured limitation.

### 5.2.2. Architectural and training-run limitations

Although the dense and MoE systems share the same attention, convolution, and decoder components, the comparison does not isolate sparse routing alone. In the MoE architecture, the first Macaron feed-forward sublayer is removed, and the second is replaced by a sparse MoE layer. The observed differences may therefore reflect the combined effects of input-conditional routing, increased stored parameter capacity, and the modified block topology. A topology-matched dense control would be required to attribute the performance differences specifically to routing. Furthermore, each configuration was trained using a single optimisation run. The speaker-level bootstrap, therefore, estimates uncertainty arising from the evaluation speakers, but not variation across random seeds or training runs.

### 5.2.3. Pathological child speech and matched controls

PAVSig is not a general-purpose ASR corpus. To the best of the author’s knowledge, it was the only available Polish child speech dataset at the time of writing, so it was used as a pragmatic evaluation set despite four limitations: it contains isolated words rather than continuous speech, has narrow lexical coverage, includes a small number of speakers, and was designed to study sigmatism rather than general speech recognition. These four properties make the Polish results difficult to interpret in two ways. First, the WER results cannot be read as typical Polish child ASR performance, since they combine ordinary child-speech acoustics with the effects of isolated-word recording, restricted vocabulary, and atypical (sigmatic) sibilant production, and the four cannot be separated from one another. Second, the routing results carry the same ambiguity: elevated divergence for PAVSig could reflect speaker age, as intended, but could equally reflect the isolated-word task format or the pathological sibilant realisations, and the analysis cannot distinguish between these explanations. PAVSig is therefore useful as a difficult Polish child-speech mismatch condition, showing how the model behaves on speech that is far from its training distribution, but it is not a matched test set for typical Polish child speech, and the Polish results should be read with this in mind throughout.

### 5.2.4. Performance disparity metric

This thesis uses the adult-child WER gap introduced in Section 3.8.2 as the main performance disparity metric. This corresponds to a group-to-norm difference, with adult speech as the reference group and child speech as the evaluated group [62]. The metric is useful because it directly matches the research questions: RQ1 and RQ2 ask whether the MoE encoder reduces the performance gap between adult and child speech. Expressing this gap in percentage points also makes the English and multilingual experiments easy to compare.

However, the gap is only one view of disparity. As Patel et al. argue, average error rates and simple bias differences should not be used alone, because they can hide variation across speaker groups and can lead to different conclusions depending on the chosen reference group or metric [62]. In this thesis, the adult-child gap summarises each group as a single corpus-level average. The child corpora span broad age ranges, but the metric produces one number per corpus. A model that reduces the average gap could still perform worse for the youngest speakers, for older children, or for age groups that are underrepresented in the corpus. The improvements reported for English and, to a lesser extent, Dutch should therefore be read as corpus-level reductions, not as evidence that all child speakers benefit equally.

The metric also does not distinguish between two different ways of reducing the gap: improving child WER while preserving adult WER, or degrading adult WER until the two groups become more similar. For this reason, the gap is always reported together with adult and child WER. This distinction is important for interpreting the results. In English, the MoE encoder reduces the gap while improving both adult and child WER, so the reduction reflects a genuine performance improvement for child speech rather than a trade-off against adult speech. In German, by contrast, **ML-MoE<sub>LB-off</sub>** improves adult WER but worsens child WER relative to the dense baseline, so the gap increases. The gap is therefore informative only when the two underlying WER values are inspected as well.

A further limitation is that the thesis uses an absolute difference in percentage points, which makes

the result easy to interpret, but it can understate proportional differences when the adult WER is low and overstate comparability across languages whose baseline WERs differ. A relative group-to-norm measure could provide a complementary view of the same disparity [62]. It was not used as the main metric here because the research questions focus on whether the absolute adult-child WER gap decreases under the MoE architecture, but relative differences would be useful in future work, especially for comparing languages with different baseline difficulty.

The adult-child WER gap should therefore be understood as a practical disparity proxy, not as a complete bias evaluation. A fuller analysis would report per-age-bin WER, speaker-level variability, worst-group performance, and possibly relative bias measures in addition to the absolute gap [4], [62]. The routing analysis in RQ3 complements the WER-based metric by showing whether the model develops age-related routing structure, but it does not replace a more granular bias evaluation. Together, the WER gap and routing analysis show that the MoE encoder can reduce the average adult-child disparity in English, while leaving open how evenly that improvement is distributed across individual child speakers and age groups.

### 5.3. Social implications

The results show that the MoE encoder reduces the adult-child WER gap for English child speakers, and the routing analysis supports the interpretation that this reflects genuine acoustic adaptation rather than a corpus artefact. This matters beyond benchmark performance. Children who depend on voice interfaces in education, healthcare, or assistive technology are directly affected by recognition errors, since these can limit access to information, participation, and support [11]. If a system works reliably for adults but fails more often for children, it reproduces an access barrier for a group that already has less control over the technologies used around them. Child ASR performance is therefore not only a technical robustness issue, but also a matter of access.

At the same time, the remaining gap is still substantial: even in the best-performing English model, child users continue to face higher error rates than adult users. Aggregate benchmarks would hide this disparity if adult and child speech were not evaluated separately [4]. The English result should therefore be read as a partial improvement, not as evidence that the access problem has been solved.

The Stage 1 result also has an implication beyond age. **EN-MoE<sub>LS</sub>** showed a smaller adult-child gap than **EN-Dense<sub>LS</sub>** despite neither model having seen child speech audio during training, which shows that input-conditional routing can generalise to an acoustically distinct, underrepresented speaker group without group-specific training data. If this property holds more generally, sparse routing could plausibly help other underrepresented speaker groups for whom labelled training data is similarly scarce, such as elderly speakers, who face comparable ASR bias for related, age-linked acoustic reasons [4]. This thesis did not test that generalisation, so it remains a hypothesis rather than a finding; evaluating whether the same routing mechanism extends to elderly speech or other underrepresented groups without group-specific fine-tuning data is left for future work (Section 5.4).

The German result is an important caution against over-generalising this hypothesis: the benefit observed for English does not transfer automatically across languages. A child-adapted MoE encoder could improve access for child speakers in one language while worsening it for child speakers in another. Deploying such a system without language-specific demographic evaluation would risk making it less fair for some children while appearing improved on average. Evaluating adult and child WER across all supported languages, not only the development language, is therefore necessary before deployment.

More broadly, the data limitations described in Section 5.2.1 are themselves part of the bias problem. The small size and narrow design of available child speech corpora reflect a wider pattern in which underrepresented groups generate less training data, which leads to weaker ASR performance, which may reduce their engagement with voice technologies and further limit future data collection [4]. MoE fine-tuning is one tool for improving performance under limited data, but it does not solve the underlying data gap. Progress on age-inclusive ASR, therefore, requires both architectural improvements and sustained investment in representative, ethically collected child speech data.

## 5.4. Future Work

**Evaluation on atypical and elderly speech.** The results suggest that MoE routing can help when speaker groups differ acoustically from the adult training norm. This may be useful beyond child speech, especially for atypical speech conditions where acoustic variability is high. The most direct extension is elderly speech: like child speech, it can differ from middle-aged adult speech in pitch, spectral shape, articulation, and speaking rate. Evaluating elderly speakers would test whether the same routing mechanism generalises from the adult-child axis to broader age-related acoustic variation.

**Mitigating bias against atypical speech.** Several methods besides MoE routing could help reduce bias against child speech and other forms of atypical speech. Data augmentation is one option. Methods such as pitch modification, spectral or formant modification, and voice conversion can generate additional child-like training examples when real child speech is limited [72], [73], [74]. However, their effectiveness depends on the quality of the generated speech, since conversion artefacts or inaccurate acoustic transformations may limit improvements in child ASR [74], [75].

Another possible direction is to make the MoE routing more controlled. In this thesis, expert specialisation emerged without demographic supervision. Future work could test whether weakly guided or constrained routing improves adaptation to atypical speech. For example, the model could be encouraged to route child speech through a subset of experts during training, so that these experts learn child-specific acoustic and phonological patterns more clearly. This would allow the model to steer specialisation instead of relying only on the router to discover it automatically. Such an approach would need to be tested carefully, because overly strict routing could reduce flexibility or harm adult speech performance.

**Data collection.** All corpora used in this thesis involve some form of elicitation, and the evaluation may not fully reflect how children speak in everyday situations. Future data collection should use a participatory design approach, involving children and families in the recording protocol, to improve speaker diversity and reduce the current over-representation of school-age children from majority-language settings. Corpora for pathological speech should include matched typically developing speakers recorded under the same protocol, so that disorder-specific effects can be isolated; the PAVSig results in this thesis cannot be interpreted this way because no such control group exists for Polish. Standardised metadata, including precise age, dialect, and disorder status, would make it easier to compare developmental effects across datasets. Community-based collection models, with appropriate consent and privacy protections for child participants, could help build larger and more diverse corpora over time.

**Bias evaluation.** The gap reduction found for English is a corpus-level average over a wide age range. Testing whether it is consistent across age subgroups within each corpus would directly strengthen or qualify the English finding. The German result also illustrates that the corpus domain interacts with the biased outcome. Intersectional evaluation combining age with language background, dialect, or speech impairment would help identify which speaker profiles benefit most from MoE adaptation and where further work is needed.

**Routing interpretability.** The JSD analysis identifies where in the encoder age-related specialisation is strongest, but does not show which acoustic features drive those routing decisions. Probing individual expert outputs for properties such as fundamental frequency, formant trajectories, or speaking rate would connect the routing results to interpretable acoustic dimensions and help explain why specialisation concentrates at the observed encoder layers.

**Developmental generalisation.** The child-adapted models were evaluated on the age ranges represented in each corpus. Whether MoE routing adapts to child speech more broadly, including age groups not seen during training or adolescents who fall between child and adult acoustic norms, remains an open question. Evaluating on a wider developmental range would test whether the gap reduction found here reflects a general adaptation to child acoustics or is specific to the acoustic profile of the training corpora.

## 5.5. Contributions

This thesis makes four main contributions.

First, it implements and releases a sparse MoE variant of the FastConformer-CTC encoder in NVIDIA NeMo, with English and multilingual configurations and an optional load-balancing auxiliary loss. The accompanying GitHub repository provides the code, training configurations, preprocessing scripts, evaluation scripts, and routing-analysis tools. The trained dense and MoE checkpoints are released separately on Hugging Face.

Code: <https://github.com/a-sasin/inclusive-asr-moe>

Model checkpoints: <https://huggingface.co/TUDELFT/inclusive-asr-moe>

Second, an empirical answer to RQ1: under a two-stage training protocol, the MoE encoder reduces the English adult-child WER gap relative to a dense baseline at both stages, without degrading adult performance, and independently of the load-balancing setting.

Third, an empirical answer to RQ2: extending the comparison to English, Dutch, German, and Polish shows that the gap reduction is language-dependent rather than a general multilingual advantage, with German as a counter-example where the gap widens.

Fourth, a routing analysis method and its findings for RQ3: per-layer routing divergence, expert preference deltas, functional similarity between expert outputs, and a domain control together show that age-related expert specialisation emerges without demographic labels and corresponds to genuinely distinct expert computations.

# 6

## Conclusion

This thesis studied whether sparse Mixture-of-Experts layers can reduce the adult-child word error rate gap in English and multilingual ASR, and whether MoE routing develops age-related expert specialisation without demographic labels.

**RQ1: English adult-child gap.** MoE reduces the adult-child WER gap at both stages of the training protocol, with the gap reduction already present before any child speech audio is used for fine-tuning. Child-speech fine-tuning further reduces the gap for both architectures, but MoE retains a significant advantage throughout, supporting an architectural contribution beyond the effect of child-speech data alone.

**RQ2: Multilingual generalisation.** The gap reduction is statistically significant only for English in the multilingual setting. German is the clearest counter-example: MoE improves adult WER but widens the adult-child gap, which may reflect domain mismatch between the available German child corpus and the adult training data, showing that the benefit of MoE depends on data quality and domain coverage as well as architecture.

**RQ3: Implicit routing specialisation.** After child-speech fine-tuning, routing divergence between adult and child speech concentrates in two regions of the encoder, one in the earlier layers and one in the later layers. Expert preference and functional divergence analyses confirm that these routing differences reflect genuinely distinct computations, and a domain-control experiment indicates that the tested recording-domain contrast does not explain them. In the multilingual model, language identity produces substantially stronger routing divergence than speaker age, but age-related routing remains detectable within individual languages.

Child speech recognition has lagged behind adult recognition for decades, driven by acoustic and articulatory differences between child and adult speech and by the scarcity of available training data [6], [10]. Prior work has addressed this gap mainly through adaptation, fine-tuning, or data augmentation [4], [6]. MoE architectures were introduced as input-conditional models that increase capacity without proportionally increasing computation [44], [45], while more recent ASR work has examined structured expert routing and robustness [19], [29]. This thesis connects these two lines of research by asking whether MoE routing itself can learn age-related structure.

The central finding is that a shared encoder, without explicit speaker-age labels, develops age-correlated expert preferences with measurable functional consequences. This extends previous work on language-conditioned routing in multilingual ASR [19], [23] and emergent language-dependent routing in multilingual MoE models [66] to a within-language speaker-age dimension, and provides evidence that age-correlated specialisation can emerge without explicit demographic supervision.

This has practical implications for inclusive speech recognition. Under suitable data conditions, a single MoE encoder can serve both adult and child speakers while reducing the performance gap relative to a dense architecture, without requiring separate model branches or explicit age labels at inference time. The routing analysis provides an interpretable account of how this happens, locating the specialisation in specific encoder layers and expert pairs. These findings point toward sparse routing as a promising and interpretable mechanism for building ASR systems that are more inclusive across speaker groups.

# References

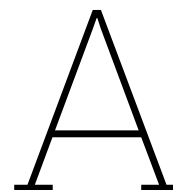
- [1] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, ser. ICML '06, New York, NY, USA: Association for Computing Machinery, 2006, pp. 369–376. DOI: 10.1145/1143844.1143891. [Online]. Available: <https://doi.org/10.1145/1143844.1143891>.
- [2] A. Gulati et al., "Conformer: Convolution-augmented Transformer for Speech Recognition," en, in *Interspeech 2020*, ISCA, Oct. 2020, pp. 5036–5040. DOI: 10.21437/Interspeech.2020-3015. [Online]. Available: [https://www.isca-archive.org/interspeech\\_2020/gulati20\\_interspeech.html](https://www.isca-archive.org/interspeech_2020/gulati20_interspeech.html).
- [3] D. Rekesh et al., "Fast Conformer With Linearly Scalable Attention For Efficient Speech Recognition," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec. 2023, pp. 1–8. DOI: 10.1109/ASRU57964.2023.10389701. [Online]. Available: <https://ieeexplore.ieee.org/document/10389701>.
- [4] S. Feng, B. M. Halpern, O. Kudina, and O. Scharenborg, "Towards inclusive automatic speech recognition," *Computer Speech & Language*, vol. 84, 2024, Art. no. 101567. DOI: 10.1016/j.csl.2023.101567. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0885230823000864>.
- [5] A. A. Attia, J. Liu, W. Ai, D. Demszky, and C. Espy-Wilson, "Kid-Whisper: Towards Bridging the Performance Gap in Automatic Speech Recognition for Children VS. Adults," en, *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 7, pp. 74–80, Oct. 2024. DOI: 10.1609/aies.v7i1.31618. [Online]. Available: <https://ojs.aaai.org/index.php/AIES/article/view/31618>.
- [6] V. Bhardwaj et al., "Automatic Speech Recognition (ASR) Systems for Children: A Systematic Literature Review," en, *Applied Sciences*, vol. 12, no. 9, p. 4419, Jan. 2022. DOI: 10.3390/app12094419. [Online]. Available: <https://www.mdpi.com/2076-3417/12/9/4419>.
- [7] Y. Zhang, Y. Zhang, B. Halpern, T. Patel, and O. Scharenborg, "Mitigating bias against non-native accents," en, in *Interspeech 2022*, ISCA, Sep. 2022, pp. 3168–3172. DOI: 10.21437/Interspeech.2022-836. [Online]. Available: [https://www.isca-archive.org/interspeech\\_2022/zhang22n\\_interspeech.html](https://www.isca-archive.org/interspeech_2022/zhang22n_interspeech.html).
- [8] E. Hermann and M. Magimai-Doss, "Handling Acoustic Variation in Dysarthric Speech Recognition Systems Through Model Combination," en, in *Interspeech 2021*, ISCA, Aug. 2021, pp. 4788–4792. DOI: 10.21437/Interspeech.2021-2212. [Online]. Available: [https://www.isca-archive.org/interspeech\\_2021/hermann21\\_interspeech.html](https://www.isca-archive.org/interspeech_2021/hermann21_interspeech.html).
- [9] C. Cucchiari, H. Van hamme, O. van Herwijnen, and F. Smits, "JASMIN-CGN: Extension of the Spoken Dutch Corpus with Speech of Elderly People, Children and Non-natives in the Human-Machine Interaction Modality," in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, N. Calzolari et al., Eds., Genoa, Italy: European Language Resources Association (ELRA), May 2006. [Online]. Available: <https://aclanthology.org/L06-1141/>.
- [10] P. G. Shivakumar and P. Georgiou, *Transfer Learning from Adult to Children for Speech Recognition: Evaluation, Analysis and Recommendations*, arXiv:1805.03322 [eess], May 2018. DOI: 10.48550/arXiv.1805.03322. [Online]. Available: <http://arxiv.org/abs/1805.03322>.
- [11] A. Koenecke et al., "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, Apr. 2020. DOI: 10.1073/pnas.1915768117. [Online]. Available: <https://www.pnas.org/doi/10.1073/pnas.1915768117>.

- [12] S. D'Arcy and M. Russell, "A comparison of human and computer recognition accuracy for children's speech," in *Interspeech 2005*, 2005, pp. 2197–2200. DOI: 10.21437/Interspeech.2005-697. [Online]. Available: [https://www.isca-archive.org/interspeech\\_2005/darcy05\\_interspeech.html](https://www.isca-archive.org/interspeech_2005/darcy05_interspeech.html).
- [13] H. Vorperian and R. Kent, "Vowel Acoustic Space Development in Children: A Synthesis of Acoustic and Anatomic Data," *Journal of Speech, Language, and Hearing Research*, vol. 50, pp. 1510–1545, Dec. 2007. DOI: 10.1044/1092-4388(2007/104).
- [14] M. Gerosa, D. Giuliani, S. 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, 7:1–7:8. DOI: 10.1145/1640377.1640384. [Online]. Available: <https://doi.org/10.1145/1640377.1640384>.
- [15] B. Łukaszewicz, "The dynamical landscape: Phonological acquisition and the phonology–phonetics link," *Phonology*, vol. 38, pp. 81–121, Jul. 2021. DOI: 10.1017/S0952675721000051.
- [16] I. S. B. Nip, J. R. Green, and D. B. Marx, "Early speech motor development: Cognitive and linguistic considerations," eng, *Journal of Communication Disorders*, vol. 42, no. 4, pp. 286–298, 2009. DOI: 10.1016/j.jcomdis.2009.03.008.
- [17] L. Rumberg et al., "kidsTALC: A Corpus of 3- to 11-year-old German Children's Connected Natural Speech," en, in *Interspeech 2022*, ISCA, Sep. 2022, pp. 5160–5164. DOI: 10.21437/Interspeech.2022-330. [Online]. Available: [https://www.isca-archive.org/interspeech\\_2022/rumberg22\\_interspeech.html](https://www.isca-archive.org/interspeech_2022/rumberg22_interspeech.html).
- [18] M. Krecichwost, Z. Miodonska, A. Sage, J. Trzaskalik, E. Kwasniok, and P. Badura, "Polish multichannel audio-visual child speech dataset with double-expert sigmatism diagnosis," *Scientific Data*, vol. 12, p. 1612, Oct. 2025. DOI: 10.1038/s41597-025-05896-8.
- [19] W. Wang, G. Ma, Y. Li, and B. Du, *Language-Routing Mixture of Experts for Multilingual and Code-Switching Speech Recognition*, arXiv:2307.05956 [cs.SD], Jul. 2023. DOI: 10.48550/arXiv.2307.05956. [Online]. Available: <http://arxiv.org/abs/2307.05956>.
- [20] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive Mixtures of Local Experts," *Neural Computation*, vol. 3, no. 1, pp. 79–87, Mar. 1991, eprint: <https://direct.mit.edu/neco/article-pdf/3/1/79/812104/neco.1991.3.1.79.pdf>. DOI: 10.1162/neco.1991.3.1.79. [Online]. Available: <https://doi.org/10.1162/neco.1991.3.1.79>.
- [21] S. Masoudnia and R. Ebrahimpour, "Mixture of experts: A literature survey," en, *Artificial Intelligence Review*, vol. 42, no. 2, pp. 275–293, Aug. 2014. DOI: 10.1007/s10462-012-9338-y. [Online]. Available: <https://doi.org/10.1007/s10462-012-9338-y>.
- [22] K. Hu, B. Li, T. N. Sainath, Y. Zhang, and F. Beaufays, *Mixture-of-Expert Conformer for Streaming Multilingual ASR*, arXiv:2305.15663 [cs.CL], May 2023. DOI: 10.48550/arXiv.2305.15663. [Online]. Available: <http://arxiv.org/abs/2305.15663>.
- [23] Y. Kwon and S.-W. Chung, "MoLE : Mixture Of Language Experts For Multi-Lingual Automatic Speech Recognition," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10096227. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/10096227>.
- [24] F. Peng, R. A. Jacobs, and M. A. Tanner, "Bayesian Inference in Mixtures-of-Experts and Hierarchical Mixtures-of-Experts Models with an Application to Speech Recognition," *Journal of the American Statistical Association*, vol. 91, no. 435, pp. 953–960, Sep. 1996. DOI: 10.1080/01621459.1996.10476965. [Online]. Available: <https://www.tandfonline.com/doi/abs/10.1080/01621459.1996.10476965>.
- [25] Z. You, S. Feng, D. Su, and D. Yu, "SpeechMoE: Scaling to Large Acoustic Models with Dynamic Routing Mixture of Experts," in *Interspeech 2021*, 2021, pp. 2077–2081. DOI: 10.21437/Interspeech.2021-478.
- [26] N. Gaur et al., "Mixture of Informed Experts for Multilingual Speech Recognition," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021, pp. 6234–6238. DOI: 10.1109/ICASSP39728.2021.9414379. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/9414379>.

- [27] M. Perez, Z. Aldeneh, and E. M. Provost, "Aphasic Speech Recognition using a Mixture of Speech Intelligibility Experts," in *Interspeech 2020*, ISCA, 2020, pp. 4986–4990. DOI: 10.21437/Interspeech.2020-2049.
- [28] Q. Zhao, G. Sun, C. Zhang, M. Xu, and T. F. Zheng, *SAML: Speaker Adaptive Mixture of LoRA Experts for End-to-End ASR*, arXiv:2406.19706 [cs], Jun. 2024. DOI: 10.48550/arXiv.2406.19706. [Online]. Available: <http://arxiv.org/abs/2406.19706>.
- [29] Z. Gu, T. Likhomanenko, and N. Jaitly, *Omni-Router: Sharing Routing Decisions in Sparse Mixture-of-Experts for Speech Recognition*, en, arXiv:2507.05724 [cs.CL], Jul. 2025. DOI: 10.48550/arXiv.2507.05724. [Online]. Available: <https://arxiv.org/abs/2507.05724v3>.
- [30] M. A. Anusuya and S. K. Katti, *Speech Recognition by Machine, A Review*, arXiv:1001.2267 [cs], Jan. 2010. DOI: 10.48550/arXiv.1001.2267. [Online]. Available: <http://arxiv.org/abs/1001.2267>.
- [31] S. Young, *Statistical Modeling in Continuous Speech Recognition (CSR)(Invited Talk)*, Jan. 2013. DOI: 10.48550/arXiv.1301.2318. [Online]. Available: <http://arxiv.org/abs/1301.2318>.
- [32] S. Alharbi et al., "Automatic Speech Recognition: Systematic Literature Review," *IEEE Access*, vol. 9, pp. 131 858–131 876, 2021. DOI: 10.1109/ACCESS.2021.3112535.
- [33] 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, Aug. 1980. DOI: 10.1109/TASSP.1980.1163420. [Online]. Available: <https://ieeexplore.ieee.org/document/1163420>.
- [34] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," en, *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989. DOI: 10.1109/5.18626.
- [35] M. Gales and S. Young, "The Application of Hidden Markov Models in Speech Recognition," en, *Foundations and Trends® in Signal Processing*, vol. 1, no. 3, pp. 195–304, 2008. DOI: 10.1561/20000000004. [Online]. Available: <http://www.nowpublishers.com/article/Details/SIG-004>.
- [36] G. Hinton et al., "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," en, *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012. DOI: 10.1109/MSP.2012.2205597.
- [37] A. Graves, A.-r. Mohamed, and G. Hinton, *Speech Recognition with Deep Recurrent Neural Networks*, arXiv:1303.5778 [cs], Mar. 2013. DOI: 10.48550/arXiv.1303.5778. [Online]. Available: <http://arxiv.org/abs/1303.5778>.
- [38] A. Hannun et al., *Deep Speech: Scaling up end-to-end speech recognition*, arXiv:1412.5567 [cs], Dec. 2014. DOI: 10.48550/arXiv.1412.5567. [Online]. Available: <http://arxiv.org/abs/1412.5567>.
- [39] A. Graves and N. Jaitly, "Towards End-To-End Speech Recognition with Recurrent Neural Networks," en, in *Proceedings of the 31st International Conference on Machine Learning*, PMLR, Jun. 2014, pp. 1764–1772. [Online]. Available: <https://proceedings.mlr.press/v32/graves14.html>.
- [40] Y. Belinkov and J. Glass, *Analyzing Hidden Representations in End-to-End Automatic Speech Recognition Systems*, arXiv:1709.04482 [cs], Sep. 2017. DOI: 10.48550/arXiv.1709.04482. [Online]. Available: <http://arxiv.org/abs/1709.04482>.
- [41] R. Sennrich, B. Haddow, and A. Birch, *Neural Machine Translation of Rare Words with Subword Units*, arXiv:1508.07909 [cs], Jun. 2016. DOI: 10.48550/arXiv.1508.07909. [Online]. Available: <http://arxiv.org/abs/1508.07909>.
- [42] A. Vaswani et al., *Attention Is All You Need*, arXiv:1706.03762 [cs], Jun. 2017. DOI: 10.48550/arXiv.1706.03762. [Online]. Available: <http://arxiv.org/abs/1706.03762>.
- [43] M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm," *Neural Computation*, vol. 6, no. 2, pp. 181–214, 1994. DOI: 10.1162/NECO.1994.6.2.181.
- [44] W. Fedus, B. Zoph, and N. Shazeer, *Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity*, arXiv:2101.03961 [cs.LG], Jun. 2022. DOI: 10.48550/arXiv.2101.03961. [Online]. Available: <http://arxiv.org/abs/2101.03961>.

- [45] N. Shazeer et al., *Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer*, arXiv:1701.06538 [cs.LG], Jan. 2017. DOI: 10.48550/arXiv.1701.06538. [Online]. Available: <http://arxiv.org/abs/1701.06538>.
- [46] A. Tjandra et al., “Massively Multilingual ASR on 70 Languages: Tokenization, Architecture, and Generalization Capabilities,” in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, ISSN: 2379-190X, Jun. 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10094667. [Online]. Available: <https://ieeexplore.ieee.org/document/10094667/>.
- [47] G. I. Winata, G. Wang, C. Xiong, and S. Hoi, “Adapt-and-Adjust: Overcoming the Long-Tail Problem of Multilingual Speech Recognition,” en, in *Interspeech 2021*, ISCA, Aug. 2021, pp. 2451–2455. DOI: 10.21437/Interspeech.2021-1390. [Online]. Available: [https://www.isca-archive.org/interspeech\\_2021/winata21\\_interspeech.html](https://www.isca-archive.org/interspeech_2021/winata21_interspeech.html).
- [48] A. Waters, N. Gaur, P. Haghani, P. Moreno, and Z. Qu, “Leveraging Language ID in Multilingual End-to-End Speech Recognition,” in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec. 2019, pp. 928–935. DOI: 10.1109/ASRU46091.2019.9003870. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/9003870>.
- [49] P. Želasko et al., *EMMeTT: Efficient Multimodal Machine Translation Training*, en, Sep. 2024. DOI: 10.48550/arXiv.2409.13523. [Online]. Available: <https://arxiv.org/abs/2409.13523v1>.
- [50] Q.-A. Dang and C. Ngo, *Polyglot-Lion: Efficient Multilingual ASR for Singapore via Balanced Fine-Tuning of Qwen3-ASR*, en, arXiv:2603.16184 [cs.CL], Mar. 2026. DOI: 10.48550/arXiv.2603.16184. [Online]. Available: <https://arxiv.org/abs/2603.16184v1>.
- [51] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “LibriSpeech: An ASR corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, Australia, 2015, pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964.
- [52] R. Ardila et al., “Common Voice: A Massively-Multilingual Speech Corpus,” in *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*, European Language Resources Association (ELRA), 2020, pp. 4218–4222. DOI: 10.48550/arXiv.1912.06670. [Online]. Available: <http://arxiv.org/abs/1912.06670>.
- [53] X. Li, S. Takamichi, T. Saeki, W. Chen, S. Shiota, and S. Watanabe, *YODAS: Youtube-Oriented Dataset for Audio and Speech*, arXiv:2406.00899 [cs.CL], Jun. 2024. DOI: 10.48550/arXiv.2406.00899. [Online]. Available: <http://arxiv.org/abs/2406.00899>.
- [54] S. S. Pradhan, R. A. Cole, and W. H. Ward, *My Science Tutor (MyST) – A Large Corpus of Children’s Conversational Speech*, arXiv:2309.13347 [cs.CL], Sep. 2023. DOI: 10.48550/arXiv.2309.13347. [Online]. Available: <http://arxiv.org/abs/2309.13347>.
- [55] D. S. Park et al., “SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition,” in *Interspeech 2019*, arXiv:1904.08779 [eess.AS], Sep. 2019, pp. 2613–2617. DOI: 10.21437/Interspeech.2019-2680. [Online]. Available: <http://arxiv.org/abs/1904.08779>.
- [56] Y. Shibata et al., “Byte Pair Encoding: A Text Compression Scheme That Accelerates Pattern Matching,” Department of Informatics, Kyushu University, Tech. Rep. DOI-TR-161, 1999.
- [57] A. Graves, *Sequence Transduction with Recurrent Neural Networks*, arXiv:1211.3711 [cs], Nov. 2012. DOI: 10.48550/arXiv.1211.3711. [Online]. Available: <http://arxiv.org/abs/1211.3711>.
- [58] O. Kuchaiev et al., *NeMo: A toolkit for building AI applications using Neural Modules*, en, arXiv:1909.09577 [cs.LG], 2019. DOI: 10.48550/arXiv.1909.09577. [Online]. Available: <http://arxiv.org/abs/1909.09577>.
- [59] K. He, X. Zhang, S. Ren, and J. Sun, *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification*, arXiv:1502.01852 [cs.CV], Feb. 2015. DOI: 10.48550/arXiv.1502.01852. [Online]. Available: <http://arxiv.org/abs/1502.01852>.
- [60] I. Loshchilov and F. Hutter, *Decoupled weight decay regularization*, 2019. arXiv: 1711.05101 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/1711.05101>.

- [61] D. P. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization*, arXiv:1412.6980 [cs.LG], Jan. 2017. DOI: 10.48550/arXiv.1412.6980. [Online]. Available: <http://arxiv.org/abs/1412.6980>.
- [62] T. Patel, W. Hutiri, A. Y. Ding, and O. Scharenborg, *How to Evaluate Automatic Speech Recognition: Comparing Different Performance and Bias Measures*, arXiv:2507.05885 [cs.CL], Jul. 2025. DOI: 10.48550/arXiv.2507.05885. [Online]. Available: <http://arxiv.org/abs/2507.05885>.
- [63] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in ASR performance evaluation," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, May 2004, pp. 409–412. DOI: 10.1109/ICASSP.2004.1326009.
- [64] Z. Liu and F. Peng, *Statistical Testing on ASR Performance via Blockwise Bootstrap*, arXiv:1912.09508 [stat.ML], May 2020. DOI: 10.48550/arXiv.1912.09508. [Online]. Available: <http://arxiv.org/abs/1912.09508>.
- [65] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, Jan. 1995. DOI: 10.1111/j.2517-6161.1995.tb02031.x. [Online]. Available: <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>.
- [66] L. Bandarkar, C. Yang, M. Fayyaz, J. Hu, and N. Peng, *Multilingual routing in mixture-of-experts*, arXiv:2510.04694 [cs.CL], 2025. DOI: 10.48550/arXiv.2510.04694. arXiv: 2510.04694 [cs.CL]. [Online]. Available: <https://arxiv.org/abs/2510.04694>.
- [67] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, Jan. 1991. DOI: 10.1109/18.61115. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/61115>.
- [68] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
- [69] L. Li et al., *Sub-MoE: Efficient Mixture-of-Expert LLMs Compression via Subspace Expert Merging*, arXiv:2506.23266 [cs.LG], Jun. 2025. DOI: 10.48550/arXiv.2506.23266. [Online]. Available: <http://arxiv.org/abs/2506.23266>.
- [70] 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:2206.07931 [eess.AS], Jun. 2022. DOI: 10.48550/arXiv.2206.07931. [Online]. Available: <http://arxiv.org/abs/2206.07931>.
- [71] A. Q. Jiang et al., *Mixtral of Experts*, arXiv:2401.04088 [cs.LG], 2024. DOI: 10.48550/arXiv.2401.04088. [Online]. Available: <http://arxiv.org/abs/2401.04088>.
- [72] V. P. Singh, H. Sailor, S. Bhattacharya, and A. Pandey, "Spectral Modification Based Data Augmentation For Improving End-to-End ASR For Children's Speech," en, in *Interspeech 2022*, ISCA, Sep. 2022, pp. 3213–3217. DOI: 10.21437/Interspeech.2022-11343. [Online]. Available: [https://www.isca-archive.org/interspeech\\_2022/singh22b\\_interspeech.html](https://www.isca-archive.org/interspeech_2022/singh22b_interspeech.html).
- [73] S. Shahnawazuddin, N. Adiga, K. Kumar, A. Poddar, and W. Ahmad, "Voice Conversion Based Data Augmentation to Improve Children's Speech Recognition in Limited Data Scenario," en, in *Interspeech 2020*, ISCA, Oct. 2020, pp. 4382–4386. DOI: 10.21437/Interspeech.2020-1112. [Online]. Available: [https://www.isca-archive.org/interspeech\\_2020/shahnawazuddin20\\_interspeech.html](https://www.isca-archive.org/interspeech_2020/shahnawazuddin20_interspeech.html).
- [74] S. Zhao, M. Singh, A. Woubie, and R. Karhila, "Data augmentation for children ASR and child-adult speaker classification using voice conversion methods," en, in *Interspeech 2023*, ISCA, 2023, pp. 4593–4597. DOI: 10.21437/Interspeech.2023-702.
- [75] W. Liu, J. Li, and T. Lee, *An Investigation on Applying Acoustic Feature Conversion to ASR of Adult and Child Speech*, arXiv:2205.12477 [eess.AS], May 2022. DOI: 10.48550/arXiv.2205.12477. [Online]. Available: <http://arxiv.org/abs/2205.12477>.



# Acknowledgements

## A.1. Acknowledgements

I would like to thank NVIDIA for supporting this work through computational resources. The access to GPU resources made it possible to train and evaluate the models used in this thesis, and was essential for carrying out the experimental part of the project.

I am deeply grateful to Prof. dr. Odette Scharenborg for her supervision, guidance, and support throughout this thesis. Her feedback helped me shape the research direction, improve the structure of the work, and think more critically about inclusive speech recognition.

I would also like to sincerely thank Dr. Piotr Żelasko for his daily supervision, technical guidance, and many helpful discussions. His support was especially valuable during the implementation and experimental stages of the project, and helped me better understand both the modelling choices and the practical challenges of working with ASR systems.

I am very thankful to both for their regular meetings, detailed feedback, and encouragement during the thesis process. Their guidance helped me develop this project from an initial research idea into a complete thesis.

I would also like to thank my colleagues for reading this thesis and offering their feedback and suggestions, which helped me improve its accuracy and clarity.

## A.2. Acknowledgement of AI Tool Usage

During this thesis, I used AI-based tools as support tools in the research, coding, and writing. These tools were not used to replace my own research work, analysis, or conclusions.

For the literature review, I used ResearchRabbit to explore relevant papers and connections between sources. I also used AI tools to summarise papers, extract key information, and simplify difficult concepts when needed. The original sources were used for the thesis content and references.

For coding, I used AI tools for debugging, implementation support, code feedback, and improving code readability. The research ideas, experiments, and methodological choices were my own. AI support was mainly used when technical problems occurred or when the code needed clearer comments for future reviewers.

For writing, I used Grammarly to check grammar, spelling, clarity, and readability. These tools helped improve the presentation of the text, but the content, research direction, and interpretation of results were my own.

The cover page of this thesis was also created with support from Canva AI, a generative AI tool.