applied sciences

MDPI

Article

Improving End-to-End Models for Children's Speech Recognition

Tanvina Patel * and Odette Scharenborg

Multimedia Computing Group, Delft University of Technology (TU Delft), 2628 XE Delft, The Netherlands; o.e.scharenborg@tudelft.nl

* Correspondence: t.b.patel@tudelft.nl

Abstract: Children's Speech Recognition (CSR) is a challenging task due to the high variability in children's speech patterns and limited amount of available annotated children's speech data. We aim to improve CSR in the often-occurring scenario that no children's speech data is available for training the Automatic Speech Recognition (ASR) systems. Traditionally, Vocal Tract Length Normalization (VTLN) has been widely used in hybrid ASR systems to address acoustic mismatch and variability in children's speech when training models on adults' speech. Meanwhile, End-to-End (E2E) systems often use data augmentation methods to create child-like speech from adults' speech. For adult speech-trained ASRs, we investigate the effectiveness of augmentation methods; speed perturbations and spectral augmentation, along with VTLN, in an E2E framework for the CSR task, comparing these across Dutch, German, and Mandarin. We applied VTLN at different stages (training/test) of the ASR and conducted age and gender analyses. Our experiments showed highly similar patterns across the languages: Speed Perturbations and Spectral Augmentation yield significant performance improvements, while VTLN provided further improvements while maintaining recognition performance on adults' speech (depending on when it is applied). Additionally, VTLN showed performance improvement for both male and female speakers and was particularly effective for younger children.

Keywords: children's speech recognition; speed perturbations; spectral augmentation; vocal tract length normalization; end-to-end automatic speech recognition

Citation: Patel, T.; Scharenborg, O. Improving End-to-End Models for Children's Speech Recognition. *Appl. Sci.* **2024**, *14*, 2353. https://doi.org/10.3390/app14062353

Academic Editors: Ying Shen, Cunhang Fan and Ya Li

Received: 29 November 2023
Revised: 23 February 2024
Accepted: 25 February 2024
Published: 11 March 2024

1. Introduction

Speech and Language Technology (SLT) solutions for children are useful for several applications, e.g., conversational interfaces for various applications, technologies for diagnosis and the treatment of a variety of developmental disorders, and in education and learning [1,2]. However, research and development on children's speech-driven SLT applications are lagging behind those for adults' speech. There are several reasons for this. Children's speech is known to be different from adults' speech in many aspects, including acoustic, prosodic, lexical, morphosyntactic, and pragmatic aspects, which are caused by physiological differences (e.g., shorter vocal tract lengths), cognitive/developmental differences (e.g., different stages of language acquisition), and behavioral differences [3–5]. For instance, children's speech exhibits increased magnitude and variability of temporal and spectral parameters in vowels, such as duration, fundamental frequency (F0), and formants (F1–F3), compared to adults' speech. When around 15 years of age, children's speech starts to resemble that of adults [6], which implies that over the course of the years children's speech changes and increasingly becomes more 'adult-like'. Moreover, the vocal tract of a child is not just a smaller version of an adult vocal tract [7]. Since acoustic features that are used for speech processing, such as the Mel-frequency Cepstral Coefficients (MFCCs) [8], are based on a model of the adult vocal tract, acoustic features might not capture the underlying child vocal tract well. Another major reason for SLT applications and research on children's speech being less well developed than those for adults' speech is the limited availability of children's speech datasets. This scarcity is partly due to stricter privacy standards associated with collecting and sharing children's data [9,10]. The shortage of

annotated children's speech data makes it a low-resource problem. These issues make the research and development of Children's Speech Recognition (CSR) systems challenging.

In this study, our main aim is to improve the performance of Automatic Speech Recognition (ASR) systems for children's speech in the absence of any children's training data (speech and text)—a situation occurring for many languages in the world and leading to large performance gaps between adults' and children's speech recognition performance. For instance, the authors of [11] showed that an End-to-End (E2E) transformer-based ASR System (without a Language Model (LM)) trained on English adults' speech of Librispeech achieved a 2.89% Word Error Rate (WER) for adults, but the WER increased to 38.8% on the MyST corpus and reached 87.2% on the OGI Kids Corpus. These high error rates are dependent on different factors, including age and speech types, and are likely to be similar for other languages. We aim to improve children's speech recognition performance by tackling the two biggest problems outlined above: the mismatch between adults' and children's speech, i.e., the variability in children's speech on one hand, and data scarcity on the other hand.

The mismatch between adults' and children's speech and variability in children's speech is often addressed by capturing the acoustic variability in children's speech through improved acoustic features (see also Section 2.2). For example, Vocal Tract Length Normalization (VTLN) [12] has been widely used to reduce the acoustic feature mismatch between adults' and children's speech due to vocal tract length variations [13,14]. Children typically have a shorter vocal tract length, resulting in higher-frequency sounds; the VTLN technique normalizes speech features based on the estimated warping factors, which account for the variations in vocal tract length. To tackle the data scarcity issue, typically, adults' speech training data are acoustically modified to resemble children's speech (see also Section 2.2), for instance, through pitch modification [15], spectral modification [16], voice conversion [17], and speed perturbations (SP) [18], and the additional, modified speech is used as (additional) training material. The chosen scenario, that no child data are available, also means that no LM will be used. Note that although LM integration could enhance performance, especially in cases of read speech by adult speakers [19], it might not effectively model the unique patterns found in children's speech as the grammar and structure of children's speech is different from that of adults. This was also observed in [11], where it was observed that an E2E transformer-based ASR trained on the adults' speech from Librispeech, without an LM, outperformed an E2E ASR with an LM incorporated when tested on children's speech from the MyST corpus and OGI Kids corpus.

In this study, we focus on E2E models for the recognition performance advantages they provide over hybrid models [20] and investigate well-known approaches in hybrid modeling for their potential in E2E children speech recognition. We compare the effectiveness of VTLN and two specifically chosen data augmentation techniques: Speed Perturbations (SP) [21] and Spectral Augmentation (SpecAug) [22]. Speed perturbation is chosen as it allows for mimicking a child's higher pitch and slower speaking rate compared to adults: increasing the speed rate of adults' speech increases the pitch and lowering the speed mimics the slower speaking style of children. Speed perturbation might thus mimic parts of children's speech. Spectral augmentation was chosen as it has often been found to make the ASR system more robust to non-read speech [23]. Research has shown that not all data augmentation techniques work for all types of diverse speech [23]; we therefore, use these commonly used augmentation approach in E2E systems, such as SP, rather than any specific pitch or frequency modification to investigate the effect of a common augmentation approach on diverse speech to further understand the limitations and possibilities of existing data augmentation techniques on diverse speech. We study the effect of augmentations (SP and SpecAug) and normalization (VTLN) separately and together. The augmentation and normalization techniques are evaluated on both children's and adults' speech, with the goal of improving children's speech recognition performance while maintaining performance on adults' speech when adapting the model to children's speech.

Summarizing, in this work our contributions are: (1) We assess the effectiveness and language independence of the data augmentation and VTLN approaches within E2E systems for three distinct languages: Dutch and German, two closely related Germanic languages, and Mandarin, an unrelated Asian tone language. (2) We analyze the effects of data augmentation and VTLN for different child age groups and gender categories. (3) Previous work on E2E models of adults' speech recognition showed that a VTLN filter-bank front-end provides better performance than the original filter-bank features [24]. VTLN's potential benefits in the context of E2E children's speech recognition are explored here for the first time. (4) Where typically VTLN models are trained on the same adult or children's speech data as the ASR model is trained on, we explore various types of training data for VTLN model training. Moreover, we assess the warping factors and effectiveness of using adult and/or children's speech as VTLN training data, as well as monolingual versus multilingual, multi-speaker speech data for VTLN training. (5) VTLN can be applied during training and testing or during testing alone, with potentially different results [25]. We investigate the effect of applying VTLN during training and testing and only during testing for E2E children's speech recognition across different languages and different speech styles. Our work can thus be considered a baseline or benchmark in E2E modeling using VTLN for children speech recognition in the absence of children's speech data for acoustic model and language model training as it provides comprehensive results and comparisons across different languages and age groups.

2. Background on Children's Speech Recognition (CSR)

2.1. Children's Speech Databases

The development of children SLT is crucially dependent on the availability of children's speech databases. Table 1 is a list of commonly used children's speech databases. A foremost difference amongst the databases is the speaker age covered. Article 1 of the UN Convention on the Rights of the Child "Definition of a child" [26] defines the "child as a human being who is below the age of 18 years". However, children's speech recognition performance is heavily influenced by the speaker's age due to the large differences in pronunciation and language use between different developmental stages in language and speech acquisition, with speech of younger children typically recognized as much worse by an ASR than that of older children and with speech from teenagers being recognized only as slightly worse than that of adults [6,27]. Age-related development stages can be broadly categorized as: newborn (ages 0–4 weeks); infant (ages 4 weeks–1 year); toddler (ages 12–36 months); preschooler (ages 3–5 years); school-aged child (ages 6–12 years); and adolescent (ages 13–19) [28]. As newborns, infants, and toddlers are beyond the scope of this work, they are not further addressed nor discussed. Most children's speech databases contain English speech [3,29–32], with only a few datasets available in other languages [33–38]. For the English language, the earlier databases (first few rows in Table 1) consisted primarily of read speech consisting of isolated words, commands, and phrases [3,29,39]. Current data collection efforts are more focused on spontaneous and conversational speech, which is however not easily obtained with children [40]. Moreover, where earlier databases consist of speech of only a small number of child speakers, current efforts focus on obtaining speech from many different speakers, which makes these databases more useful for the development of children's speech technology. Recognizing native children's speech thus already presents inherent challenges. These challenges increase when recognizing non-native children's speech. This is due to the non-native accents in the speech, which is caused by the influence of the native language on the pronunciation of the words in the non-native language, mispronounced words, ungrammatical utterances, disfluencies (false starts, partial words, and filled pauses), and code-switched words. Several databases containing non-native accented (children's) speech have been created that can be used to develop speech technology for non-native (child) speaker groups [34,41]. The availability of these databases in the public domain along with the organization of several challenges [42–44] has sped up the much-needed development and improvement of CSR systems. In this

work, we use the SLT CSR Mandarin speech data [37], kidsTALC German speech data [38], and a part of the Jasmin Dutch [36] corpus. All these corpora are available on request from the respective sources for research purposes, which allows for easy benchmarking and comparisons of the approaches and systems.

Table 1. Common children's speech databases with details of the language, type of speech, age range, and number of speakers (#Spk).

Databases	Language	Type of Speech	Age Range	#Spk
CID, 1996 [39]	en	Read speech	5–18 years	436
SAIL-Inhouse, 1997 [3]	en	Digit, commands and phrases	5–18 years+ Adult	-
CMU Kids Corpus, 1997 [29]	en	Read speech	6–11 years	76
CHIMP, 1998 [45]	en	Read speech	6–14 years	97
OGI Kids Corpus, 2000 [30]	en	Spontaneous speech	KG to 10th Grade	1100
TBALL, 2005 [46]	en, es	Read speech	5–8 years	256
PF STAR, 2005 [34]	en, de, sv, it	Read and spontaneous (native + non-native)	4–15 years	158
FBK ChildIt, 2005 [35]	it	Read speech	7–13 years	170
CU Read, 2006 [31]	en	Isolated words and sentences	6–11 years (1–5 Grade)	663
CU Story, 2006 [32]	en	Read and summarized stories	8–11 years (3–5 Grade)	106
CGN-Jasmin, 2006 [36]	nl	Read and machine interaction speech (native+non-native)	6–18 years	190
My Science Tutor (MyST), 2011 [40]	en	Children and virtual tutor conversations	7–11 years (3–4 Grade)	1370
TLT-school, 2020 [41]	en, de	Spontaneous speech: Non-native	9–16 years	10k
SLT CSR Challenge, 2021 [37]	zh	Read and conversational	4–11 years + Adult	980
kidsTALC, 2022 [38]	de	Continuous speech	3.5–11 years	47

KG = Kindergarten, Languages: ISO 639-1 language codes: English = en, Spanish = es, German = de, Swedish = sv, Italian = it, Chinese = zh, and Dutch = nl.

2.2. Brief Overview of Research on Children's Speech Recognition Using Adult-Speech-Based ASRs

Given the limited availability of children's speech data, many studies employ ASR models trained on and for adults' speech to recognize children's speech, as illustrated in Figure 1 (where solid lines indicate training and dashed lines indicate testing). Although using ASR models trained on adults' speech for the recognition of adults' speech in many cases leads to reasonably good results, using these for children's speech recognition results in sub-optimal recognition performance due to the acoustic and linguistic differences in adults' and children's speech (Figure 1a).

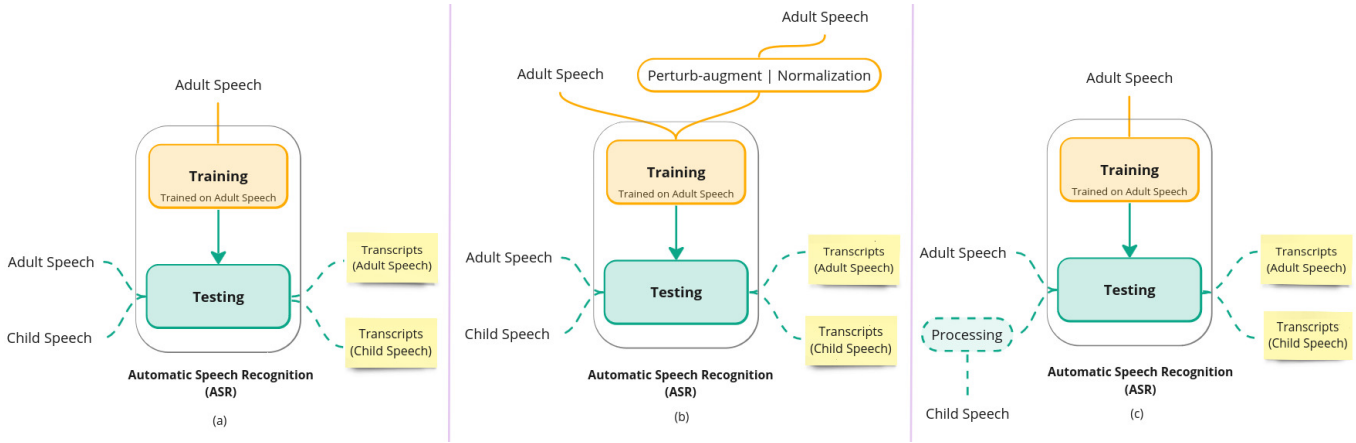

Figure 1. (**a**) ASR model trained on adults' speech and tested on adults' and children's speech; (**b**) modification of adults' speech or features during the training phase to make the adults' speech more children's speech-like (solid orange lines); (**c**) modification of the children's speech or features during the test phase to make the children's speech more adults' speech-like (green dashed block). Solid lines indicate training and dashed lines indicate testing. Approach (**b**,**c**) can also be combined, like in this paper.

To improve children's speech recognition, various approaches have been explored to adapt adult speech-based systems. These approaches can be categorized into two main strategies, indicated in Figure 1b,c. Firstly, Figure 1b: Adaptation of *adult training* data: adults' speech is perturbed, e.g., its speed, duration, or pitch is acoustically modified to resemble children's speech, or the spectral features are perturbed through spectral feature modifications [16], adaptation techniques [4], speaker adaptive training [13], or normalization techniques [12]. The modified adult data is then added to the pre-existing adults' speech, and the ASR system is retrained and used for testing of the adults' and children's speech (green, dashed lines). Secondly, Figure 1c: Modification of children's speech or acoustic features *during test:* The ASR is only trained on adults' speech, as in Figure 1a. However, the children's speech is modified prior to testing to match the characteristics of the adults' speech or features used in training.

Until recently, hybrid Acoustic Model and Language Model (AM-LM) systems dominated CSR applications and focused on capturing the acoustic variability in children's speech through improved acoustic features [15,47]. More recently, E2E models have been proposed for CSR applications in which, typically, adults' speech training data is acoustically modified to resemble children's speech using augmentation or conversion techniques [17,48]. In the remainder of this section, we present a brief (non-exhaustive) overview of research on CSR covering the commonly used approaches where perturbed adults' speech is used for training, first for hybrid systems and then for E2E systems. Note that some of the techniques included in this overview assume the availability of children's speech for training/fine-tuning (unlike the scenario in our research). We describe the databases and languages used; the models and techniques used, along with the training and test details; and the improvements obtained with the applied technique, focusing on the choice of training speaker groups (adult and/or child). This analysis helps us identify gaps and informs our approach to enhance children's speech recognition. An overview of the papers that use hybrid models is presented in Tables 2 and 3 for the E2E models.

Table 2. Summary of CSR research using hybrid ASR systems: databases details, techniques, and models used.

Literature	Corpus:Lang.	Techniques	Training | Test Details	Model	Metric: Improvement
Potamianos et al., 2003 [4]	SAIL-own: en	Frequency warping, model adaptation	Adult | Children	GMM	WER: 15.9% to 8.7%
			Adult + Children Children	GMM	WER: 7.6% to 5.6%
Ghai et al., 2009 [14]	TIDIGITS: en	VTLN and spectral smoothing	Adult | Children	GMM	WER: 11.3% to 2.15%
Cosi 2009 [49]	FBK ChildIt: it	SMAPLR and VTLN	Adult Female | Children	GMM	PER: 28.7% to 18.0%
			Adult + Children | Children	GMM	PER: 17.4% to 12.3%
Shahnawazuddin 2016 [47]	PF-STAR: en	Adaptive-liftering + VTLN	Adult WSJCAM0 | PF_STAR	DNN	WER: 24.2% to 21.4%
Kathania et al., 2018 [15]	PF-STAR: en	Loudness, voice-intensity, and voice-probability	Adult WSJCAM0 | PF_STAR	DNN	WER: 19.6% to 12.7%
			Adult + Children | Children	DNN	WER: 11.4% to 8.82%
Shivakumar et al., 2020 [50]	Multiple: en	Transfer learning (based on TEDLIUM)	Adult TEDLIUM—CU Kid's, OGI, CHIMP | CID	TDNN	WER: 39.3% to 7.8%
Kathania et al., 2022 [51]	PF-STAR: en	Formant modification, VTLN, and SAR	Adult WSJCAM0 | PF_STAR	TDNN	WER: 14.1% to 8.69%

Table 3. Summary of CSR research using E2E ASR systems: databases details, techniques, and models used.

Literature	Corpus:Lang.	Techniques	Training | Test Details	Model	Metric: Improvement
Chen at al. 2020 [18]	SLT Challenge: zh	Pitch, speech, tempo, volume, and reverberation	Adult + Children | Children	E2E	CER: 16.2% to 13.6%
Ng et al., 2020 [52]	SLT Challenge: zh	Transfer Learning (base on Adult)	Adult | Child	E2E	CER: 38.5% to 23.6%
		SpecAugment, RIR, and volume perturbation	Adult | Child	E2E-LM	CER: 23.6% to 20.1%
Gelin et al., 2021 [53]	Lailo: fr	Simulating reading mistakes	Adult Common Voice | Lailo	E2E	PER: 22.9% to 19.9%
Shivakumar et al., 2022 [11]	MyST,OGI Kids: en	Fine-tune on MyST | greedy decoding	Librispeech + librivox | MyST	E2E-LM	WER: 25.46% to 16.01%
Singh et al., 2022 [16]	Internal:en	Spectral warping, formant perturbation, and VTLP	Librispeech Adult | Child	E2E	WER:36.1% to 32.2%
Zhao et al., 2023 [33]	Samromur: Icelandic	Formant, pitch, and vowel stretching	Adult | Child	E2E	WER: 51.3% to 36.3%
			Adult + Child | Child	E2E	WER: 28.7% to 26.5%

Hybrid systems generally use Gaussian Mixture Models (GMM) or Deep Neural Network (DNN)-based AM-LM and are typically evaluated using Word Error Rate (WER) as the evaluation metric (see also Table 2). Typically, in hybrid systems trained with adults' speech, research focuses on developing acoustic features that capture or reduce acoustic variability in children's speech or make the adult acoustic features more like the features of children's speech, often using feature normalization that normalizes vocal tract differences (i.e.,vocal-tract-related features) or that normalizes source differences (i.e., source-level features). The following are examples of vocal-tract-related features: Work in [4] introduced front-end frequency warping and filtering techniques on digit recognition tasks; others used VTLN-based normalization to reduce acoustic variability in features used for the CSR task [13,14,47]. The VTLN features were found to work for both English and Italian [25,49] (see Table 2, second column for the codes of the languages). At the source-level, pitch modifications change adults' speech pitch frequency (source) to that of children's speech (target) [15,54], and formant modification approaches, which aim to reduce the difference in vocal tract resonances (formant frequencies) between adult and child speakers [51], have been found to enhance CSR systems. A second approach that is investigated to improve CSR is to add children's speech data to the training data for training from scratch or fine-tuning, which adapts the adult ASR to the children's speech [50]. However, children's speech data are not always available, making this approach not always feasible.

Next, we discuss several approaches used in E2E systems for CSR, as detailed in Table 3. The E2E systems require a substantial amount of children's speech training data for CSR. Hence, the commonly used approach is to acoustically modify adults' speech such that is sounds more like children's speech. For instance, in [18], the authors used various data augmentation techniques, including pitch, speed, tempo, volume perturbation, spectral augmentation, and reverberation, resulting in nearly tenfold more training data. In [53], data augmentation is carried out by synthesizing reading mistakes (simulating word-level repetitions and substitutions) to improve recognition performance of E2E models. In [16], segmental spectrum warping and perturbations in formant energy are introduced to generate a children-like speech spectrum from that of an adult's speech spectrum. Voice conversion by changing the spectral characteristics of adults' speech is also explored to generate child-like speech from adults' speech as in [17,33]. Secondly, in situations where some children's speech data are available, transfer learning approaches could be used [52] or the acoustic and language models are adapted to the children's speech and text [11]. Moreover, recently, a common trend has been to enhance recognition performance by fine-tuning large self-supervised speech models with target data, e.g., recent work in [55] demonstrated that fine-tuning Whisper on children's speech significantly improved CSR performance for several English language databases, compared to non-fine-tuned Whisper models. However, similar to transfer learning, this approach involves using children's speech, which differs from our scenario where only adults' speech is used.

Summarizing Tables 2 and 3 shows that in the scenario that no children's speech data are available for training, for hybrid systems, most research is carried out on English and read speech data, while the more recent research using E2E models focus on a larger variety of languages. Feature adaptation techniques such as VTLN for children's speech are restricted to GMM-HMM and DNN-HMM systems, which give the best performances on (English) read speech in the range of 10–20% WER. Recognition performances with E2E systems on several other languages reach similar performances on read speech and more than 30% WER on spontaneous speech. Current trends in E2E models employ data creation/generation techniques to improve CSR. However, the feasibility and effect on the recognition performance of using normalization or adaptation in E2E based models, especially for children's speech recognition, is unexplored and is the topic of this paper.

3. Methodology

This section provides an overview of the datasets (see Table 4), the augmentation and normalization techniques, the training configurations, and the experimental set-up used.

Table 4. Details of the Dutch, German, Mandarin adults' and children's speech datasets used in this study.

Language	Datasets	Speaking Style	Age Range	#Speakers	#Utterances | #Hours			
					Training	Validation	Test-Read	Test-CTS/HMI
Dutch	CGN	Read | CTS	18–65	2897	704,293 | 433	70,498 | 43	409 | 0.45	3884 | 1.80
	Jasmin-DC	Read | HMI	06–13	71	-	-	13,104 | 6.55	3955 | 1.55
	Jasmin-DT	Read | HMI	12–18	63	-	-	9061 | 4.90	2723 | 0.94
	Jasmin-NnT	Read | HMI	11–18	53	-	-	11,545 | 6.03	3093 | 1.16
German	CommonVoice	Read	-	3413	330,454 | 452.8	55,050 | 75.23	55,102 | 75.5	-
	kidsTALC	Read	3–11	47	9187 | 9.18	1676 | 1.68	2413 | 2.2	-
Mandarin	SLT-SetA	Read	18–60	1999	196,616 | 276.7	22,764 | 31.52	23,950 | 33.41	-
	SLT-SetC1	Read	7–11	927	24,109 | 23.38	2701 | 2.48	3042 | 2.79	-
	SLT-SetC2	Conversational	4–11	166	25,245 | 23.49	2955 | 2.85	-	3303 | 3.14

3.1. Databases

The Dutch Corpora: *Corpus Gesproken Nederlands (CGN)* [56]: The CGN is a corpus containing Dutch speech spoken by native speakers from the Netherlands and Flanders. In this study, we only used the data recorded in the Netherlands. The corpus consists of monologue and multilogue speech spoken by speakers within the 18–65 age range. It has 15 different speech types, which include read speech, lecture recordings, broadcast data, spontaneous conversations, and telephonic speech. The unprocessed training data consist of the speech from all components summing to around 480 h spoken by 1187 female and 1710 male speakers. Two test sets were used: broadcast news (BN) and conversational telephone speech (CTS). The pre-processing procedures and test partitions are the same as in [57]. The Dutch models used in this study are trained on the adults' speech from the CGN corpus.

Jasmin-CGN corpus [36]: Jasmin-CGN corpus is an extension of the CGN corpus. It consists of read speech and Human–Machine Interaction (HMI) speech spoken by Dutch children, teenagers and older adults, and teenagers and adult non-native speakers of Dutch. Here, all native and non-native children's and teenager's speech (Dutch Native Children: DC; Dutch Native Teenagers: DT; and Dutch Non-Native Teenagers: NnT) are used for testing only.

The German Corpora: *Mozilla's Common Voice (CV)*: The German adults' speech dataset used in this study is obtained from Mozilla's Common Voice (CV) project [58]. It is a large open-source dataset consisting of read speech where speakers contribute speech by reading words from a screen. The dataset contains metadata, including the gender, age, and accent region of the speaker. In this work, around 600 h of standard-accented German from adult speakers is used. The training, validation, and test splits are those as used in [59]. The German models are trained on the adults' speech from the CV corpus.

kidsTALC dataset: The German children's speech dataset used in this study is the kidsTALC dataset [38]. The kidsTALC dataset is specifically designed to support the development of speech-based technological solutions and contains 25 h of continuous speech from children aged 3^1/$_2$–11 years. In this study, all children's speech from the kidsTALC validation set is used for testing our models (as the kidsTALC test set does not contain transcripts).

The Mandarin Corpus: This dataset was released as part of the Children Speech Recognition Challenge (CSRC) organized as an event of the IEEE Spoken Language Technology 2021 workshop [37]. It consists of 3 sets of Mandarin speech data, i.e., Set A contains 341 h of adult read speech, Set C1 consists of 29 h of child read speech, and Set C2 of 30 h of child conversational speech. In this paper, we follow the data splits for training, validation, and test from [52]. For the Mandarin experiments, we train the Mandarin ASR with Set A of adults' speech and test on the test sets of Set A, Set C1, and Set C2.

3.2. Augmentation and Normalization Techniques

Speed Perturbation (SP): SP involves resampling the original raw speech signal, which results in a warped time signal. Given an audio speech signal $s(t)$, time warping by a factor β gives the signal $s(\beta t)$. The Fourier transform of $s(\beta t)$ is $S(\omega/\beta)/\beta$. As a result of the changes in the time domain, which affect the number of frames in the utterance, the time warping produces shifts in the frequency components (shift of the speech spectrum); thus, speed perturbation affects both tempo and pitch [21]. The adults' speech training data was perturbed at 90% and 110% of the original rate, creating a 3-fold training set.

Spectral Augmentation (SpecAug): SpecAug squeezes and stretches the spectrogram locally and has been found to improve recognition performance in conversational and non-read speech type scenarios [22]. It is applied to the log mel spectrogram of the input audio rather than the raw audio itself. It consists of three augmentation policies: (1) time masking; (2) frequency masking (that masks a block of consecutive time steps or mel frequency channels); and (3) time-warping, which randomly warps the spectrogram along the time axis. SpecAug was applied with its default settings, i.e., the maximum width of each frequency mask $F = 30$, the maximum width of each time mask $T = 40$, the number of frequency and time masks $= 2$, and the masked parts are filled with the mean.

Vocal Tract Length Normalization (VTLN): The vocal tract length varies from person to person and is quite different for children and adults. The differences in vocal tract length lead to differences in the spectrum, i.e., the formant frequencies shift in frequency. The process of compensating spectral variation due to the length of the vocal tract is known as VTLN. It is a normalization technique that is often used for speaker recognition and related tasks [60]. Basic normalization techniques linearly scale the center frequencies of the filter bank in the front-end feature extractor to approximate formant frequency scaling [61]. More recent approaches include calculating a linear feature transform for each VTLN warp factor, i.e., [12].

$$x^{\alpha} = A^{\alpha}x + B^{\alpha} = W^{\alpha}X; \quad (1)$$

where $W^{\alpha} = [A^{\alpha}; B^{\alpha}]$ is the affine transformation matrix, x is the feature vector, α is the warping factor (chosen using grid search), x^{α} is the transformed feature vector for warp factor α, A^{α} is the linear transformation matrix and B^{α} is the linear bias for warp factor α, and X is the extended feature set [62]. Once the warping factors are estimated, a piece-wise linear warping function is implemented that maps the frequency range in three segments. Let the warping function be $W(f)$, where f is the frequency. The central segment maps f to f/α, where α is the VTLN warp factor (typically in the range 0.8 to 1.2). The process of VTLN warps the features to that of an ideal or reference speaker ($\alpha_{ref} = 1$). In adult male speakers, the energy in the speech spectrum tends to be concentrated towards the lower frequencies, whereas in adult female speakers, it is generally higher; hence, their estimated warping factors are around $\alpha_{male} \geq \alpha_r$ and $\alpha_{female} \leq \alpha_r$, respectively. For children, since their spectrum energies are typically even higher than those for female speakers, it is expected that $\alpha_{child} < \alpha_r$ compresses the frequency axis closer to the reference. Feature normalization with VTLN is a two-step process, i.e.,

1. Train a VTLN model on a given speech dataset.
2. Estimate the warping factor α for an utterance and normalize its features with the factor α.

3.3. Experimental Setup

Baseline CSR System: For our experiments, we use the conformer architecture, which combines convolution neural networks and transformers to model both local and global dependencies, respectively [63,64]. We use the ESPnet toolkit [19] to run the experiments [65]. We train separate E2E systems for each of the three languages, i.e., Dutch, German, and Mandarin, using the adults' speech of the respective languages, and test the systems on their respective adults' and children's speech test sets. All of the audio files are single-

channel and recorded at a 16 kHz sampling rate. The training configuration of the baseline E2E model uses 80 dimensional log-mel filterbank features with 3-dimensional pitch features. The experiments were carried out till 20 epochs. For training the Dutch and German ASR systems, we use byte pair models with 5000 and 1000 unigram tokens, respectively. Since Mandarin is not an alphabetic language, the Mandarin ASR system instead uses a dictionary with ∼6 k characters. The non-language symbols (speaker filler, laugh, and unknown) are only available for the Dutch language and are thus only used in the Dutch experiments. The Mandarin and German databases mostly consist of read speech and are likely not to contain fillers (which are also not annotated). To evaluate the performance of our models, we compare them to a state-of-the-art pre-trained models, the Open-AI Whisper models [66]. Detailed results are provided in Appendix A. In short, for adults' speech, our model outperformed Whisper small, medium, and large despite not using a language model, for all three languages, except for German, where Whisper large outperformed our model, and Mandarin where Whisper medium outperformed our model. Hence, we carry out our experiments with the conformer model.

Augmentation and Normalization Experiments: To study the effect of augmentation and normalization, three experiments were carried out:

Experiment 1: Augmentation—First, the adults' speech of each of the three languages was perturbed using SP. For each language, the perturbed adults' speech, combined with the original adults' speech, was then used to retrain the respective baseline models, after which these "SP-augmented" models were retrained with SpecAug applied during training (SP + SpecAug). The SP and SP + SpecAug models were tested on the children's speech of the respective languages; note that no changes were made to the audio signal of the test sets. This scenario is depicted in Figure 1b.

Experiment 2: Normalization—The Effect of VTLN: For each language, we train a VTLN model on only the adults' speech. VTLN training does not require any transcriptions of the speech data. Since children's speech data without transcriptions is (more) often readily available (than with transcriptions that are often challenging to obtain), we add an experiment where the VTLN models are trained on (untranscribed) children's speech. Table 5 provides an overview of the trained VTLN models and the type of data they were trained on. VTLN is implemented during the feature extraction process as follows: during the training process of the ASR, the adults' speech training set features are normalized using the estimated VTLN factors obtained with the VTLN model trained on adults' speech or the VTLN model trained on children's speech, as indicated in Table 5. ASR models (following the baseline set-up) are then trained with these normalized features, one new ASR model for each VTLN model (i.e., three for Dutch and two for both German and Mandarin). This training scenario is depicted in Figure 1b. During test, the features of the adults' and children's speech test sets are extracted and normalized to avoid any mismatch between the training and testing acoustic features and subsequently passed through the ASR system for decoding. This testing scenario is depicted in Figure 1c.

Table 5. Overview of the different VTLN models trained on the three languages; the age ranges of the speakers; and the types of speech.

Language	VTLN Model	VTLN Training Dataset	Age Range	Speech Type
Dutch	$VTLN_{CGN}$	CGN-train	18–65	Read
	$VTLN_{Jas\text{-}DCDT}$	Jasmin-{DC, DT}	6–18	Read + HMI
	$VTLN_{Jas\text{-}DCDTNnT}$	Jasmin-{DC, DT, and NnT}	6–18	Read + HMI
German	$VTLN_{CV}$	CV-train	-	Read
	$VTLN_{kidsTALC}$	kidsTALC-train	3–11	Read + Conversational
Mandarin	$VTLN_{SetA}$	SetA-train	18–60	Read
	$VTLN_{SetC1C2}$	SetC1C2-train,dev	4–11	Read + Conversational
Multiple	$VTLN_{MultiV1}$	Randomly selected ∼5 h of data from different speaker groups in Dutch, German, and Mandarin	3–65+	Read + HMI + Conversational

Experiment 3: The Combined Effect of Augmentation and Normalization: To study the effect of augmentation and normalization combined (i.e., the combination of the scenarios in Figure 1b,c), we combine the set-ups of Experiment 1 (the ASR models are trained with SP + SpecAug) and Experiment 2, with the difference that VTLN is applied to the ASR models in two ways:

- *During training and testing*: Each of the VTLN models for each language is applied during training to adult training data as in Figure 1b and also applied during testing to the adults' and children's speech test sets as in Figure 1c, yielding three ASR models for Dutch and two ASR models for both German and Mandarin.
- *During testing only*: Each of the VTLN models for each language is only applied during the test stage to the adults' and children's speech test sets as in Figure 1c.

Moreover, in addition to the language-specific adults' and children's speech-based VTLN models, we also trained a VTLN model ($VTLN_{MultiV1}$) with speech data from diverse speaker groups from all three languages. Specifically, we randomly selected around 5 h of speech from the Dutch, German, and Mandarin adults' and children's speech databases, selecting from the three available speaking styles (read, HMI, and conversational speech). The motivation for training a VTLN model on such diverse speech is that preliminary findings have shown that VTLN models can be applied cross-lingually [67]. We therefore explore the feasibility of using a common VTLN model that could work across languages.

Evaluation: Recognition performance for the Dutch and German experiments is reported in WER and for Mandarin in Character Error Rate (CER). WER and CER are calculated as the ratio of word/character insertion, substitution, and deletion errors in the recognized transcription and the total number of spoken words/characters in the ground truth transcription [68]. We conduct an analysis of the estimated warping factors for each speaker group in the adults' and children's test sets to investigate the link between the estimated warping factors and the recognition performance. To further understand the possibilities and limitations of VTLN, we analyzed the WER results with respect to age and gender for Dutch (as this information is only available in the Dutch-Jasmin children's speech database). To assess the statistical significance of the results, we employ the procedure outlined in [69], which utilizes matched pairs sentence-segment word error (MAPSSWE) [70] to determine whether the observed differences in WER are statistically significant. We report p-values (provided in Appendix B) to indicate significant differences at the levels of $p = 0.001$ (*), $p = 0.01$ ("‡"), or $p = 0.05$ ("†").

4. Results and Analyses

4.1. Baseline Model Performance

Dutch: The results of the Dutch experiments are presented in Table 6 split out for Dutch adults' speech (CGN), Dutch Children (DC), Dutch Teenagers (DT) and Dutch Non-Native Teenagers (NnT), for read and HMI speech, separately. Moreover, the averages over the three children/teenager speaker groups for both speaking styles are reported. The system trained on Dutch adults' speech without any augmentation or normalization, i.e., the baseline model in Table 6 (row a), achieves a 9.6% WER and 23.9% WER on read and CTS adults' speech on the CGN test sets, respectively. These results compare well to those reported in the literature on the CGN datasets, where [71] obtained a WER of 6.6% and 21.6% on CGN read and CTS speech, respectively, with a TDNNF and RNNLM model, (i.e., with the use of an LM, which we do not use in E2E models). To further evaluate our baseline model, we also trained a TDNN-BiLSTM model with SP and a tri-gram LM, which achieved WERs of 7.0% and 26.4% on CGN read and CTS speech, respectively, which are similar to those of [71] while not using an RNNLM. This indicates that our Dutch conformer E2E model is a strong baseline model. On children's speech, the baseline model shows, as expected, much worse results: an average of almost 40% WER for read speech and 50% for HMI speech, with the best results for DT and the worst for the NnT. On the same database, but with the use of an RNNLM, the authors of [27,71] reported somewhat better results: an average of 27.3% WER for read speech and 31.7% for HMI speech. Our TDNN-BiLSTM

model obtained a WER of 40.2% for read speech and 47.7% for HMI speech, which is a bit worse than the results by [71]. The difference can be attributed to the lack of an RNNLM for our TDNN-BiLSTM model, which improved the performance of the teenagers speech in [71].

Table 6. Results in %WER, with significance levels, for the Dutch ASR when trained on CGN adults' speech and tested on CGN adult and the native and non-native children's and teenager's speech from Jasmin-CGN, split for read speech and conversational speech. The lowest WERs for each speaker group across all systems are highlighted in bold.

			CGN		**Jasmin: Read**			**Jasmin: HMI**			**Jasmin: Avg**	
Training	**Augmentation**	**Normalization**	**Rd**	**CTS**	**DC**	**DT**	**NnT**	**DC**	**DT**	**NnT**	**Read**	**HMI**
CGN:Adult	(a) None	None	9.6	23.9	42.9	22.1	54.0	50.2	40.1	59.9	39.7	50.1
	(b) SP	None	7.0 *	22.0 *	36.7 *	20.5 *	55.6	43.8 *	35.4 *	60.3 †	37.6	46.5
	(c) SP + SpecAug	None	**7.0** *	**20.2** *	**36.1** *	**18.8** *	**51.1** *	**40.1** *	**27.8** *	**52.6** *	**35.3**	**40.2**
	(d) None	$VTLN_{CGN}$	9.3	23.6	38.8 *	21.2 *	53.4	45.9 *	34.9 *	59.0	37.8	46.6
	(e) None	$VTLN_{Jas\text{-}DCDT}$	9.3	24.1	36.3 *	21.8	54.1	42.0 *	35.5 *	58.6 †	37.4	45.4
	(f) None	$VTLN_{Jas\text{-}DCDTNnT}$	9.5	24.2	**35.0** *	21.2 ‡	53.0 ‡	41.1 *	32.8 *	57.5 *	36.4	43.8

† $p < 0.05$, ‡ $p < 0.01$, and * $p < 0.001$.

German: The results of the German experiments are presented in Table 7. The results are reported for adults' speech from the CV dataset and for children's speech from the kidsTALC dataset. The baseline model, in Table 7 (row a), achieves a WER $< 10\%$ on German adults' speech, showing that this is a fairly strong baseline, similar to prior work using the same dataset, which achieved a WER of 5.8% using a Conformer-RNN-T model with SpecAug augmentation [59]. However, the baseline model performs drastically worse on children's speech. The performance on the kidsTALC dataset is close to 78% WER. The only available results on the kidsTALC dataset used models trained on both kidsTALC and CV [38,72]. The work in [38] reported a 32.5% PER (WER not reported) when trained with kidsTALC and CV, while [72] reported a 21.5% PER (47.8% WER) with Wav2Vec fine-tuned on kidsTALC and CV. To further evaluate our baseline model, we trained our conformer model (without LM) on the kidsTALC and CV databases, with SP and SpecAug, which achieved a 5.1% WER on the adult CV speech and 40.6% WER on the kidsTALC development set. Thus, our conformer baseline is a good enough baseline. The performance drop from adult to children's speech can be attributed to the high acoustic variability in children's speech and partly to the difference in the recording conditions of the datasets.

Table 7. Results in %WER, with significance levels, for the German ASR when trained on CV adult database and tested on CV adults' and children's speech from the kidsTALC dataset. The lowest WERs for each speaker group across all systems are highlighted in bold.

Training	**Augmentation**	**Normalization**	**CV**	**kidsTALC**
CV: Adult	(a) None	None	9.6	77.9
	(b) SP	None	6.7 *	71.1 *
	(c) SP + SpecAug	None	**5.1** *	**67.2** *
	(d) None	$VTLN_{CV}$	9.7	76.4 ‡
	(e) None	$VTLN_{kidsTALC}$	9.8 *	72.3 *

† $p < 0.05$, ‡ $p < 0.01$ and * $p < 0.001$.

Mandarin: The results of the Mandarin experiments are shown in Table 8. The results are reported on the adults' speech of Set A, and the two children's speech datasets Set C1 (read) and Set C2 (conversational) and average over both children's speech sets. The baseline model, in Table 8 (row a), obtains a CER of 16% on the read adults' speech and

an average CER of 33.5% on the children's speech test sets, which is better than the best reported results so far of 38.5% averaged over Set C1 and Set C2 [52], indicating that we have a strong baseline. Interestingly, for the read children's speech, the results are highly similar to those for the adult read speech; however, the performance drops dramatically for the conversational children's speech to a CER of 50.8%. A possible reason for the degraded performance of Set C2 is two-fold: first, the difference between read and conversational speech where conversational speech is always harder to recognize than read speech; second, Set C2 contains speech from younger children than those in Set C1.

Table 8. Results in %CER, with significance levels, for the Mandarin ASR when trained on SLT adult database and tested on adults' and children's speech of the SLT database. The lowest CERs for each speaker group across all systems are highlighted in bold.

Training	Augmentation	Normalization	SetA	SetC1	SetC2	Average
SetA: Adult	(a) None	None	16.4	16.1	50.8	33.5
	(b) SP	None	11.0	11.2	43.6	27.4
	(c) SP + SpecAug	None	9.9 *	**10.0** *	**38.8** *	**24.4**
	(d) None	$VTLN_{SetA}$	16.3 *	15.5 *	46.4 *	31.0
	(e) None	$VTLN_{SetC1C2}$	16.7 *	15.7 *	46.3 *	31.0

† $p < 0.05$, ‡ $p < 0.01$, and * $p < 0.001$.

4.2. Experiment 1: Augmentation

Dutch: For Dutch, adding speed-perturbed native adults' speech data to the training data Table 6 (row b; see Appendix B, Table A2 for the *p*-values) led to a significant performance improvement for native children's and teenager's speech, with absolute improvements ranging from 1.6% (DT-Read) to 6.4% (DC-HMI). However, adding speed-perturbed adults' speech to the training data led to a significant performance degradation for the non-native teenagers for HMI speech. Apparently, adding more native-accented data to the training data increases the bias against non-native accented speech. This can potentially be attributed to the fact that the SP applied to CGN increases the native speech variability but not the non-native variability. Adding SpecAug (row c) led to a significant further reduction in WER for all native speaker groups, and is (as expected) particularly effective in the case of HMI speech, which showed the largest improvements. Unlike the SP-condition, the SP + SpecAug condition led to recognition results for the non-native speakers that significantly outperformed the baseline. Overall, the combined effect of SP + SpecAug gave an absolute average improvement of 7.2% over the children's speech baseline. This improvement for children's speech did not come at the cost of a deterioration for adults' speech. Rather, adding SP + SPecAug to the training data also led to a significant WER improvement of 2.6% on read adults' speech and 3.7% on continuous adults' speech.

German: For German, the effects of adding SP data and SpecAug are similar to those obtained for Dutch. Focusing on the children's speech, Table 7 (rows b, c; see Appendix B Table A3 for the *p*-values) shows that adding SP leads to a significant 6.8% absolute improvement for children's speech, while additionally adding SpecAug further increases performance with 3.9% WER. Adding perturbed adults' speech data (SP) along with SpecAug during training also led to an improvement on the adults' speech, with an overall significant improvement of 4.5% WER.

Mandarin: For Mandarin, adding SP as shown in Table 8 (row b; see Appendix B Table A4) improves children's speech recognition for both read (Set C1) and spontaneous (Set C2) speech, with an average improvement of 6.1% CER, although this improvement is not significant. The best performance is again obtained for the combination of SP and SpecAug, leading to an average improvement over the baseline of 9.1%, which significantly outperforms the baseline model for all datasets. The largest absolute improvement for both SP and SP + SpecAug (Table 8, row c) was found for the spontaneous speech in Set C2 (12% vs. 6.1% for Set C1) as compared to the baseline. This is in line with the findings for Dutch,

where SpecAug is particularly effective in the case of non-read speech compared to read. Also, we observe that adding adult perturbed data with SpecAug results in a significant performance improvement for the adults' speech (Set A) of 6.5% compared to the baseline.

4.3. Experiment 2: Normalization—The Effect of VTLN

To study the impact of VTLN, we first calculated and analyzed the warping factors for the children's speech test sets in each language. These factors were estimated using the VTLN models in Table 5. The distribution of the warping factors for each test set is visually represented in the box plots in Figure 2. In each plot, the lower and upper ends of the whiskers correspond to the minimum and maximum values of the warping factors. The straight solid orange horizontal line inside the box represents the median, while the green triangular marker indicates the mean of the warping factors. The dotted red line in each plot is the average value of the warping factors for adults' speech in that language.

Dutch: The warping factor plots are shown for the VTLN_{CGN} and the $\text{VTLN}_{\text{Jas-DCDTNnT}}$ models (note that the $\text{VTLN}_{\text{Jas-DCDT}}$ model gave almost the same warping factors as the $\text{VTLN}_{\text{Jas-DCDTNnT}}$ model). In both Dutch plots, the left three warping factors in each plot correspond to read speech (non-shaded boxes), while the right three warping factors correspond to HMI speech (shaded boxes). With the VTLN_{CGN} model, i.e., VTLN trained on only CGN adults' speech, all speaker groups have almost the same warping factors, $\alpha < 0.9$ (Figure 2a), which indicates that the warping factors have not been estimated well. This may be due to the fact that the model is trained with only adults' speech from CGN rather than also with children's speech whose warping factors we aimed to estimate. However, when the VTLN model is trained with children's speech, $\text{VTLN}_{\text{Jas-DCDTNnT}}$, the warping factors are estimated well (children's speech $\alpha < 1$ and adults' speech $\alpha \approx 1$) (Figure 2b). These warping factors are similar to those reported for other studies in English [73]. The warping factors for read and HMI speech from both Dutch VTLN models are highly similar, only slightly higher for HMI speech than for read speech.

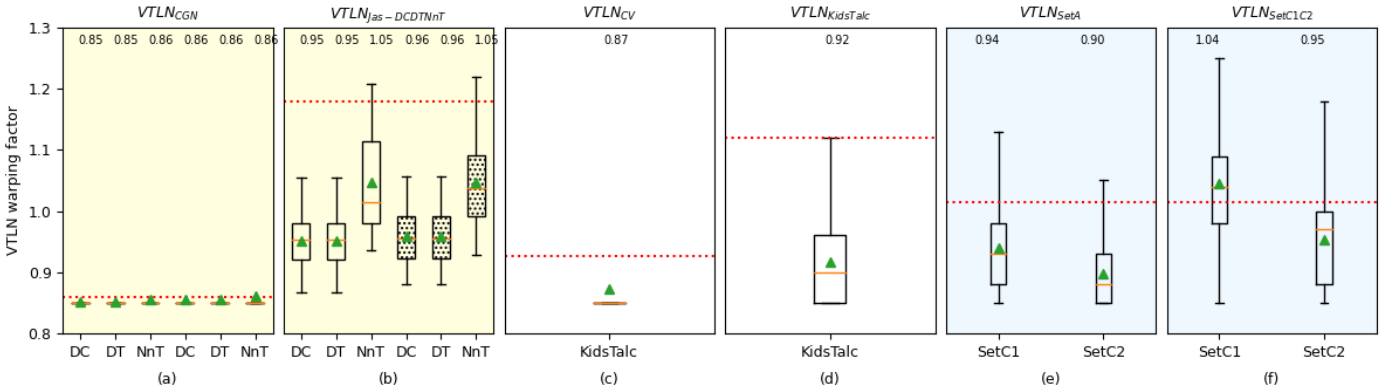

Figure 2. Box plots of the warping factors estimated across different child speaker groups using different VTLN models. Dutch: (**a**) VTLN_{CGN}, (**b**) $\text{VTLN}_{\text{Jas-DCDTNnT}}$; German: (**c**) VTLN_{CV}, (**d**) $\text{VTLN}_{\text{kidsTALC}}$; and Mandarin: (**e**) $\text{VTLN}_{\text{SetA}}$, (**f**) $\text{VTLN}_{\text{SetC1C2}}$. For each children's speech test set, the orange horizontal line within the box is the median. The triangular marker indicates the mean of the warping factors (also indicated in text at the top of the Figure). The top and bottom range represent the minimum and maximum values of the warping factors. Red dotted line: warping factor for adult speakers.

The effect of VTLN on ASR system performance without any augmentations is shown in Table 6 (rows d–f). With VTLN_{CGN}, despite the not-so-great estimation of the warping factors, the WER is lower than the baseline (row a) for almost all speaker groups but only significantly so for the native child speaker and teenager speaker groups. This is probably due to the warping factors having $\alpha < 0.9$, which means that the children's speech frequencies are lowered, making them more similar to adults' speech, hence leading to a smaller mismatch with the adult training speech and improved recognition performance. The better

estimated warping factors of $VTLN_{Jas\text{-}DCDT}$ (not shown in Figure 2) and $VTLN_{Jas\text{-}DCDTNnT}$ (see Figure 2b) led to a further improvement for children's speech (rows e, f), with the best results for the VTLN model trained on speech data that also included the non-native children's speech, yielding a significant improvement over baseline for all speaker groups. For DC-Read, the performance is even better than when SP and SpecAug are applied (compare Table 6 rows (c) and (f)). On average, the performance with $VTLN_{Jas\text{-}DCDTNnT}$ is better than $VTLN_{CGN}$. Nevertheless, even the VTLN model trained with adults' speech only provided a significant improvement over the baseline. For the adult test sets, results for the different VTLN models are not significantly different from baseline.

German: The estimated warping factors in Figure 2c,d show that $VTLN_{CV}$ led to relatively constant warping factors (indicated by the very short whiskers), whereas $VTLN_{kidsTALC}$ exhibited more varied warping factors for children's speech (much longer whisker), similar to what we observed for Dutch for $VTLN_{Jas\text{-}DCDTNnT}$: a VTLN training set with more diverse children's speech led to more diverse warping factors. The results in Table 7 (row d, e) show that the performance of the VTLN-based models significantly outperform the baseline model (Table 7, row a) for children's speech; however, they do not outperform the augmentation-based models. Similar to the Dutch findings, the use of the VTLN model trained on adults' speech only ($VTLN_{CV}$) resulted in a modest though significant improvement (1.5% absolute WER), while $VTLN_{kidsTALC}$ showed a larger improvement of 5.6% absolute WER as compared to baseline. Possibly, the varying warping factors estimated for the kidsTALC test set resulted in better normalization and performance improvement. Applying VTLN to adults' speech did not lead to improvements compared to the baseline when the VTLN model was trained on adults' speech, and in fact it significantly degraded performance when the VLTN model trained on children's speech was used.

Mandarin: Unlike what was observed for Dutch and German, the estimated warping factors in Figure 2e,f show that both $VTLN_{SetA}$, trained on adults' speech, and $VTLN_{SetC1C2}$, trained on children's speech, exhibit varying warping factors. The warping factors estimated for Set C2 (spontaneous speech) are in a lower range than Set C1 (read speech). This difference can likely be attributed to Set C2 containing speech of children younger, with shorter vocal tract length, than those in Set C1. The shorter vocal tract lengths result in higher frequencies in the speech spectrum, which require more frequency scaling (normalization) and thus lower warping factors.

Table 8 (rows d, e) shows the results of the model trained only with VTLN. Both VTLN models significantly outperform the baseline model (row a) on children's speech. The VTLN models trained on adults' speech ($VTLN_{SetA}$) and children's speech ($VTLN_{SetC1C2}$) perform similarly (row d, e), giving around 4% absolute improvement on Set C2 and approximately 0.5% absolute CER improvement on Set C1 compared to the baseline. We hypothesize that the larger improvement on Set C2 is due to the presence of speech from younger children, where we expect the normalization to be more effective based on the size of the warping factors rather than due to the conversational nature of the speech. Applying the VTLN models to the adults' speech of Set A decreased performance significantly when the VTLN model trained on children's speech was applied, though a small but significant improvement was found when the VTLN model trained on adults' speech was used.

4.4. Experiment 3: The Combined Effect of Augmentation and Normalization

Until now, we have discussed the effects of augmentation and normalization separately. The results show that overall all approaches outperform the baseline systems on children's speech, where the combination of SP and SpecAug gave the best results (except for Dutch children, DC-Read, where using normalization alone by the $VTLN_{Jas\text{-}DCDTNnT}$ model gave the best results). For Experiment 3, all of the approaches (augmentation and normalization) are combined, and we investigate different data sets for training and applying VTLN.

Dutch: In Table 9 (see Appendix B Table A5 for the *p*-values), we present the results of combining SP and SpecAug with different VTLN models when VTLN was applied during training and testing and only during testing. The results are again split for adult read (Rd)

and continuous (CTS) speech and for DC, DT, and NnT split for Read and HMI speech. The averages over the three speaker groups are also provided. For easy reference, the system trained on adults' speech with augmentations (SP and SpecAug) but without normalization is provided again (same as Table 6, row c). When applying VTLN during training and testing, we observe that all of the models significantly outperform the no-VTLN model for all three speaker groups for read speech, with $VTLN_{MultiV1}$ giving the best result for DC and DT, with improvements of 5.5% and 1.1%, respectively, and $VTLN_{Jas\text{-}DCDTNnT}$ giving the best results for the non-native teenagers, with an improvement of 1.0% WER absolute. $VTLN_{MultiV1}$ gives almost the same results on the non-native teenagers (a difference of only 0.1% absolute). Both these models are also trained on non-native accented speech, which is likely the reason for their improved results on non-native accented teenager speech. For HMI speech, applying VTLN only led to significant improvements for the native DC. No significant differences compared to baseline were found for the DT and NnT speaker groups. Applying VTLN during training and testing does not improve adults' speech.

Several general observations can be made about the results when VTLN is applied only during testing. Only for native children speech did applying VTLN lead to significant improvements over baseline for read and HMI speech, while for the native teenagers and non-native teenagers applying VTLN only during testing led to significant performance degradation compared to the baseline. For the native children, the largest improvement for read speech was 1.4% with $VTLN_{MultiV1}$, and for HMI speech 2.2% with $VTLN_{CGN}$. For adults' speech, applying VTLN during testing led to significant performance degradations for all of the models except $VTLN_{CGN}$. This shows that in scenarios where the ASR system cannot be retrained, normalizing the test data alone using VTLN warping factors can reduce the acoustic mismatch between features of adults' and children's speech and improve children's speech recognition; however, it may degrade performance for older children and adults.

Table 9. Results in %WER, with significance levels, for the Dutch ASR system trained with SP and SpecAug when VTLN was applied (1) during training and testing or (2) only during test. The lowest WERs for each speaker group are highlighted in bold for the two normalization approaches separately. Underline indicates the best result for the specific speaker group over all conditions.

			CGN		Jasmin: Read			Jasmin: HMI			Jasmin: Avg	
Training	**VTLN**	**$VTLN_{model}$**	**Rd**	**CTS**	**DC**	**DT**	**NnT**	**DC**	**DT**	**NnT**	**Read**	**HMI**
CGN adults' speech SP + SpecAug	None	None	**7.0**	**20.2**	36.1	18.8	51.1	40.1	27.8	52.6	35.3	40.2
	Train | Test	$VTLN_{CGN}$	7.3	**20.2**	34.0 *	17.9 *	50.5 *	**37.5** *	**27.4**	**52.2**	34.1	**39.0**
		$VTLN_{Jas\text{-}DCDT}$	7.6 ‡	**20.2** ‡	31.7 *	17.9 *	50.7 ‡	38.2 *	29.2	54.7	33.4	40.7
		$VTLN_{Jas\text{-}DCDTNnT}$	7.2	20.4	32.4 *	18.1 *	**50.1** *	39.0 *	29.7	54.0	33.5	40.9
		$VTLN_{MultiV1}$	7.3	20.6	**31.6** *	**17.7** *	50.2 *	38.0 *	29.2	52.7	**33.2**	39.9
	Test	$VTLN_{CGN}$	**7.0**	**20.4**	35.0 *	19.2 †	**51.8** *	**37.9** *	29.1 ‡	**52.7**	**35.3**	39.9
		$VTLN_{Jas\text{-}DCDT}$	7.9 *	21.5 *	35.3 *	19.5 *	52.7 *	39.0 ‡	28.8 ‡	55.1 *	35.8	40.9
		$VTLN_{Jas\text{-}DCDTNnT}$	7.9 *	21.6 *	35.2 *	19.4 *	52.6 *	38.9 *	28.9 ‡	54.8 *	35.7	40.8
		$VTLN_{MultiV1}$	7.7 *	20.6 †	**34.7** *	**19.1** †	52.0 *	38.2 *	**28.2**	53.2	**35.3**	**39.8**

† $p < 0.05$, ‡ $p < 0.01$, and * $p < 0.001$.

German: As shown in Table 10 (see Appendix B Table A6 for the p-values), when VTLN is applied during both training and testing, the two VTLN models for which their training data included children's speech significantly improved over baseline. The largest improvement of 2.0% is observed for the $VTLN_{kidsTALC}$ model. Slight though significant performance degradations were found when applying VTLN models to test adults' speech.

When VTLN is used only during testing, all of the models show a significant performance improvement for children's speech compared to the no-VTLN condition, with the largest improvement of 2.8% for the $VTLN_{CV}$ model. The $VTLN_{MultiV1}$ model also performed similarly to the $VTLN_{CV}$ model. Perhaps surprisingly, for children's speech,

applying VTLN only during testing outperforms the condition where VTLN is applied during training and testing for all of the models (further discussion on this is provided in Section 5). Again, slight but significant degradations were observed when applying VTLN for adults' speech.

Table 10. Results in %WER, with significance level, for the German ASR system trained with SP and SpecAug when VTLN was applied (1) during training and testing or (2) only during test. The lowest WERs for each speaker group are highlighted in bold for the two normalization approaches separately. Underline indicates the best result for the specific speaker group over all conditions.

Training	VTLN	$\text{VTLN}_{\text{model}}$	CV	KidsTALC
CV adults' speech SP + SpecAug	None	None	5.1	67.2
	Train | Test	VTLN_{CV}	**5.1**	66.5
		$\text{VTLN}_{\text{KTalc}}$	5.2 *	**65.2** *
		$\text{VTLN}_{\text{MultiV1}}$	**5.1** ‡	65.9 ‡
	Test	VTLN_{CV}	**5.2** *	**64.4** *
		$\text{VTLN}_{\text{kidsTALC}}$	5.5 *	65.0 *
		$\text{VTLN}_{\text{MultiV1}}$	**5.2** *	64.5 *

† $p < 0.05$, ‡ $p < 0.01$, and * $p < 0.001$.

Mandarin: As shown in Table 11 (see Appendix B Table A7 for the *p*-values), when VTLN is applied during both training and testing, no significant changes were observed for the children's speech of Set C1, while for Set C2, all three models significantly improved the recognition of the spontaneous children's speech compared to the no-VTLN condition. Similar to Dutch, the best performing VTLN model is $\text{VTLN}_{\text{MultiV1}}$, which gave a significant 0.9% improvement over the no-VTLN condition for Set C2. Using VTLN during training and testing did not further improve the results for Set C1, potentially because the recognition results for Set C1 were already quite good, i.e., at the same level as those for the adults' speech in Set A, leaving very little room for additional gain by applying VTLN. Recognition performance did not change much for adults' speech when applying VTLN in addition to SP and SpecAug during training and testing.

Table 11. Results in %CER, with significance levels, for the Mandarin ASR system trained with SP and SpecAug when VTLN was applied (1) during training and testing or (2) only during test. The lowest CERs for each speaker group are highlighted in bold for the two normalization approaches separately. Underline indicates the best result for the specific speaker group over all conditions.

Training	VTLN	$\text{VTLN}_{\text{model}}$	SetA	SetC1	SetC2	Average
SetA Adults' Speech SP + SpecAug	None	None	9.9	**10.0**	38.8	24.4
	Train | Test	$\text{VTLN}_{\text{SetA}}$	9.8	10.2	38.1 ‡	24.2
		$\text{VTLN}_{\text{SetC1C2}}$	9.8	10.2	38.2 ‡	24.2
		$\text{VTLN}_{\text{MultiV1}}$	**9.8**	**10.1**	**37.9** *	**24.0**
	Test	$\text{VTLN}_{\text{SetA}}$	9.9 *	10.0	37.2 *	23.6
		$\text{VTLN}_{\text{SetC1C2}}$	10.2 *	10.2	37.8 *	24.0
		$\text{VTLN}_{\text{MultiV1}}$	10.0 *	**10.0**	**37.1** *	**23.5**

† $p < 0.05$, ‡ $p < 0.01$, and * $p < 0.001$.

When VTLN was used only during testing, similar to German, further significant improvements were observed for spontaneous speech in Set C2, with again the best model being $\text{VTLN}_{\text{MultiV1}}$ which gave a 1.7% improvement over the no-VTLN condition (further discussion on this is provided in Section 5). However, again no performance differences were observed for the read speech of Set C1, while for Set A a small though significant degradation compared to baseline for the adults' speech was observed.

4.5. The Effect of VTLN: Analysis by Age

Experiment 3 showed that applying VTLN in addition to SP and SpecAug gave the best recognition performance on children's speech for all three languages; however, we also observed differences regarding these improvements based on the age groups of the child speakers for Dutch. We also saw differences for Set C1 and Set C2 for Mandarin, but we cannot disentangle the effects of younger age in Set C2 compared to Set C1 and the use of spontaneous speech in Set C2 compared to read speech in Set C1. Here, we further investigate the relationship between the effect of VTLN and the child's age for Dutch as this dataset has the largest age range and provides speaker age for most speakers, i.e., approximately 95% of the DC speaker group, 64% of the DT speaker group, and 100% of the NnT group. We use these data in our analyses. Given the overall better performance of the $\text{VTLN}_{\text{MultiV1}}$ model for all languages, we used this model for our analysis. Figure 3 shows the WER by speaker's age (in years) for the model trained with SP and SpecAug and no VTLN (blue); the model trained with SP, SpecAug, and $\text{VTLN}_{\text{MultiV1}}$ applied during training and testing (orange); and the model SP, SpecAug, and $\text{VTLN}_{\text{MultiV1}}$ only applied during testing (red), for the different speaker groups (DC, DT, and NnT) for read and HMI speech separately. From Figure 3, the following observations can be made:

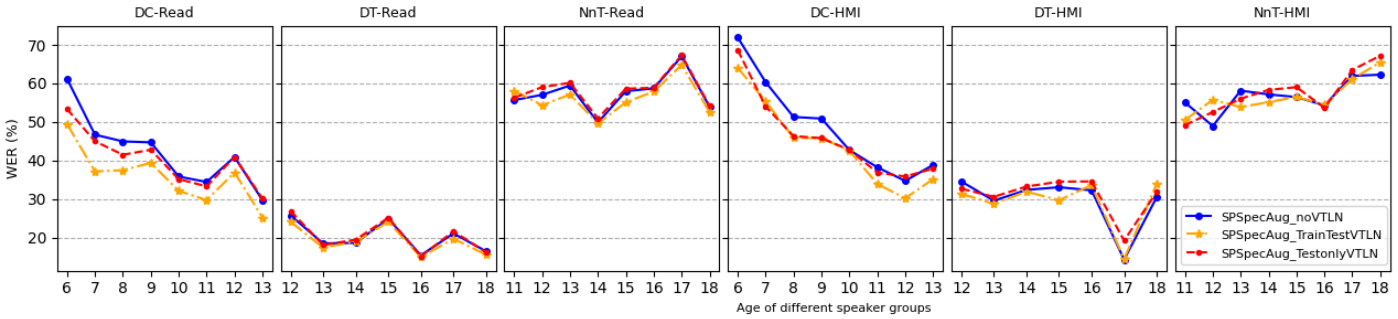

Figure 3. Age-wise average WERs for the Dutch Jasmin test sets for different speaker groups (DC, DT, and NnT) and different speech types (Read, HMI). Models: ASR model with SP and SpecAug and without any normalization (blue), ASR with $\text{VTLN}_{\text{MultiV1}}$ normalization both while training and testing (orange), and ASR with $\text{VTLN}_{\text{MultiV1}}$ normalization only while testing (red).

- Dutch native children (DC; ages 6 to 13 years): For both read and HMI speech, WERs are highest for the youngest children and progressively decrease with increasing age. This pattern is observed for all three models. Table 9 already showed that for DC, using the $\text{VTLN}_{\text{MultiV1}}$ model during training and testing gave better results than when only applying the model during testing. From Figure 3, we can see that this improvement when applying $\text{VTLN}_{\text{MultiV1}}$ is largest for the younger children for read speech, while the improvement when applying $\text{VTLN}_{\text{MultiV1}}$ is more or less the same across the different ages for HMI.
- Dutch native teenagers (DT; ages 12 to 18 years): For both read and HMI speech, similar to the DC age group, the youngest speakers have the highest WER, which progressively decreases with increasing age, although this decrease in WER is less pronounced compared to the DC speaker group. Interestingly, the pattern and WERs over the ages is highly similar for the three models. The only small effect of using VTLN during training and testing compared to the no-VTLN conditions is possibly due to the fact that as children grow, their vocal tract characteristics (especially length) are closer to those of adults and their speech spectrum has a similar frequency range to that of adults so normalization is not really needed.
- Dutch non-native teenagers (NnT; ages 12 to 18 years): The pattern observed for Dutch non-native teenagers is different from that for the Dutch native children and native teenagers. For read speech, there is no improvement in WER with increasing age but rather a small deterioration. This deterioration with increasing age is even more

pronounced for HMI speech. This is potentially explained by the fact that the older a child/person is when learning a non-native language the less likely it is to achieve native proficiency [74], so the older speakers are likely to have stronger accents. The effect of VTLN seems to be highly similar for the different ages, which is similar to the findings for the Dutch teenagers: no particular age particularly benefits more from the application of VTLN.

In summary, for native Dutch speakers, WERs decrease with age, especially in the younger age groups. Applying VTLN improves performance but is more effective for younger children, which can be explained by their higher-pitched voices, which could benefit more from vocal tract length normalization than the relatively less high-pitched voices of older children and teenagers. For non-native speakers, WERs are consistently higher and do not show a decrease but rather an increase with increasing age, with VTLN exhibiting no impact across ages. These results highlight the interplay of language proficiency and the effectiveness of different approaches at improving diverse children's speech recognition.

4.6. The Effect of VTLN: Analysis by Gender

In this section, we split the results based on gender (only two genders, male and female, are provided in the meta data, so we will use this binary split). Figure 4 displays the WERs for the female speakers (red lines) and the male (blue lines) speakers (for whom age information is available) across the different ages for the three different models and the two speaking styles. Table 12 summarizes the average WERs for both genders (over all speakers, including those for which no age information was available) per speaker group and per speaking style for the three models. From Figure 4 and Table 12, we can observe:

- Dutch native children (DC; ages 6 to 13 years): For both read and HMI speech, the earlier observed trend of decreasing WER with increasing age holds for both male and female speakers. Without any normalization (solid lines), WER is higher for females compared to males (see also Table 12). Applying VTLN has the largest effect on the female speakers: for both VTLN conditions, the WER of the female speakers is lower than that of the male speakers. The effect is largest when VTLN is applied during both training and testing (dotted lines). These results are likely due to the higher-pitched female voices compared to the male voices.
- Dutch native teenagers (DT; ages 12 to 18 years): For read speech, for all of the models female speech is recognized better than male speech. Applying VTLN seems to have a positive effect on both genders, except that the effect seems to be a bit larger for the youngest female speakers, i.e., the 12 year old, which is in line with the findings for the female child speakers in the DC group (see the left-most panel in Figure 4). For HMI speech, the same picture holds except that the effect of applying VTLN results in slightly degraded performance for the male speakers (see Table 12). In Figure 4, the fluctuations in performance observed for the 15, 16, and 17 year old female speakers are due to a small number of speakers per age (1, 6, and 1, respectively).
- Dutch non-native teenagers (NnT; ages 12 to 18 years): For the Dutch non-native Teenagers, WERs were lower for the female speakers compared to the male speakers for both read and HMI speech. This gender gap was particularly large for the more spontaneous HMI speech. Overall, applying VTLN improved recognition performance for both the female and male speakers for read speech, and approximately to the same extent (Table 12). For HMI speech, no improvement was found when applying VTLN for the male speakers, while a small improvement was found for the female speakers for the model where VTLN was applied during testing. As shown in Figure 4, this small improvement was driven by the younger female speakers, where the largest improvement was found—in line with the earlier age results.

In summary, for all speaker groups and both speech types, the average WER over all of the models was always lower for the female speakers than the male speakers (see Table 12, bottom row), which is in line with earlier findings on this data set [27]. Overall, the use of

VTLN was similar for both genders, except for the Dutch children speaker group, which showed a larger improvement for the female speakers (5.3% for read and 3.2% for HMI speech) compared to the male speakers (3.7% for read and 1.0% for HMI speech), which was largely driven by the improvement for the youngest female speakers. This gender gap can be explained by the higher-pitched voices of especially the younger female speakers, which could then benefit most from the normalization step.

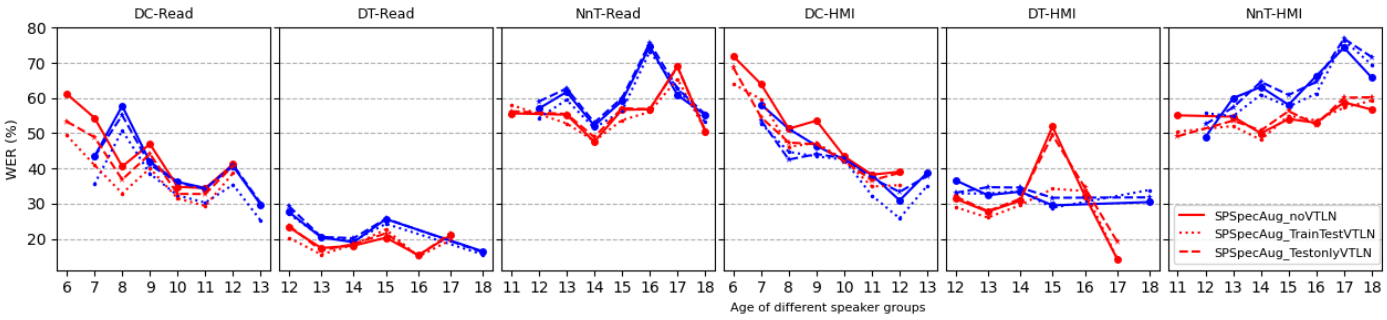

Figure 4. Age-wise average WER for female (red lines) and male (blue lines) speakers of the Dutch Jasmin test sets for different speaker groups (DC, DT, and NnT) and different speech types (Read, HMI). Models: ASR model with SP and SpecAug and without any normalization (solid), ASR with $\text{VTLN}_{\text{MultiV1}}$ normalization both while training and testing (dotted), and ASR with $\text{VTLN}_{\text{MultiV1}}$ normalization only while testing (dashed).

Table 12. Average WER for male and female speakers of the Dutch Jasmin test sets for different speaker groups (DC, DT, and NnT) and different speech types (Read, HMI) for the model with SP and SpecAug without any normalization (None), with $\text{VTLN}_{\text{MultiV1}}$ normalization both while training and testing (Train | Test), and $\text{VTLN}_{\text{MultiV1}}$ normalization only while testing (Test). The lowest WERs for each speaker group are highlighted in bold.

	DC-Read		**DT-Read**		**NnT-Read**		**DC-HMI**		**DT-HMI**		**NnT-HMI**	
Normalization	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**
None	36.26	35.92	17.18	20.39	50.32	52.01	40.67	39.44	**26.38**	**29.09**	50.56	**54.59**
Train | Test	**30.94**	**32.17**	**15.61**	**19.8**	**49.49**	**50.82**	**37.44**	**38.48**	26.81	31.46	50.67	54.75
Test	33.77	35.59	16.95	21.18	50.94	53.14	37.71	38.73	26.69	29.95	**50.42**	56.06
Average	33.66	34.56	16.58	20.46	50.25	51.99	38.61	38.88	26.63	30.17	50.55	55.13

5. General Discussion and Conclusions

In this study, we investigated data augmentation (speed perturbations and spectral augmentation) and feature normalization techniques (vocal tract length normalization) for E2E children's speech recognition in the scenario that there are no children's speech and text data available for (re)training the ASR system. We investigated the effect of these three approaches in isolation and together across three different languages, different speaking styles, and different children/teenager age groups and compared the results to those for adults' speech. For these languages, the baseline ASR models trained on adults' speech achieved a WER of <10% on adults' speech and were found to be close to or even better than state-of-the-art results for the respective data sets/languages, and they outperformed the state-of-the-art OpenAI-Whisper small and medium models (and even the large models for Dutch and Mandarin). However, performance deteriorated substantially when tested on children's speech. For Dutch, a drop of 30% absolute was observed for read speech and of over 40% absolute for the more spontaneous human–machine interaction speech. For German continuous speech, the deterioration was close to 70% absolute. For Mandarin, the picture was slightly different. Here, there was no performance drop from adult read speech to children's read speech; however, a drop of around 16% absolute (around 50% relative)

was observed for the children's speech data set, which consisted of spontaneous speech, including that of younger children than in the read speech set.

The lack of performance drop from adult to children's read speech for Mandarin is likely at least partially explained by the fact that for Mandarin, the adults' and children's speech sets are part of the same corpus, with the same recording conditions, while for Dutch and German the adults' and children's speech came from separate databases. To assess the impact of database mismatch, we tested the SLT Mandarin SP + SpecAug model (without LM) on three other Mandarin read speech databases: Magic data [75] (43 k utterances, 52 h, 78 speakers), Aishell [76] (7 k utterances, 10 h, 20 speakers), and THCHS-30 [77] (2 k utterances, 6.3 h, 10 speakers), and we obtained 2.48%, 4.10%, and 14.25% CER, respectively. The results for Magic data and Aishell are good despite database mismatch, suggesting that the relatively easy recognition task of read speech may counterbalance the effect of database mismatch, which is in line with the results for our adults' and children's Mandarin read speech. However, THCHS-30, which consists of longer utterances, shows a drop in performance. This shows that the impact of database mismatch (also) depends on the specific database characteristics. Given that for both Dutch and German the children's speech is partially or entirely non-read speech, the observed drop in recognition performance from adult to children's speech is only partially explained by the database mismatch and is thus also due to the acoustic differences between adults' and children's speech.

Similar to what has been observed before in a Mandarin E2E system for children's speech in literature [18], applying speed perturbations reduced the WERs for children's speech recognition. Performance was further improved when SpecAug was added. The beneficial effect of SpecAug is in line with findings for English children's speech, which showed an improvement when applying SpecAug over a condition without SpecAug [78]. Our results confirm and extend these earlier findings with a few observations: we observed improvements when using speed-perturbed adults' speech for children's speech recognition. We attribute this to the pitch and speed changes caused by the speed perturbations, which make the adults' speech more similar to children's speech. However, this positive effect of adding perturbed adults' speech was only observed for native speakers and was absent for Dutch non-native speakers. Applying SpecAug led to performance improvements for all speaker groups, with a more substantial impact on non-read speech types. This emphasizes that augmentation techniques may not always and uniformly enhance performance but rather depend on specific characteristics of the speaker group and speech type, which is in line with findings from [23].

As far as we are aware, we are the first to apply VTLN to adults' speech for the improvement of children's speech recognition in E2E models. Our results showed that the application of VTLN improved children's speech recognition across the board both when applying models trained on adults' speech only and when trained on children's speech (from the same database as the test data) only; however, the improvement was smaller than for the combined SP and SpecAug data augmentation methods. The combination of SP, SpecAug, and VTLN, however, gave the best children's speech recognition results for all three languages. Similar to what has been found for hybrid models [50], VTLN, even when trained on adults' speech only, thus also improves the recognition performance of children's speech in the absence of children's speech training data in E2E models without any language model. This result not only shows that VTLN provides a complementary approach and improvement to data augmentation but also that the same approach can be used across languages to improve children's speech recognition. Moreover, since we tested different types of speech (read, HMI, and spontaneous speech), these results show that the combined approach also generalizes over speech styles. Importantly, the performance on adults' speech was maintained. Thus, reducing spectral variation resulting from vocal tract length differences, which are particularly relevant to children's speech, does not impact performance on adults' speech recognition.

In our experiments, we trained different VTLN models using adults' speech and children's speech from native and non-native speakers (Dutch only) and from three different

languages. Each VTLN model exhibited variations in estimated warping factors, impacting ASR performance to varying degrees. Notably, when the VTLN model estimated warping factors that were distinct for adults and children, this generally led to improved recognition performance, particularly for younger children. In line with our no children's speech and text data scenario, we trained a VTLN model on adults' speech only, which showed significant improvements over baseline for all native children and teenager speaker groups for all three languages when applied in isolation and in combination with SP and SpecAug. Not surprisingly, training the VTLN model (also) on children's speech further improved performance. This is as expected as the VTLN model was trained on the same database as the children's speech database, thus reducing database mismatch and providing the target speech to the VTLN model for training. This scenario is nevertheless realistic as, although often both speech and transcribed text are not available for acoustic model and language model training, children's speech audio alone is more readily available. Using VTLN trained with in-domain children's speech is likely thus the best solution; however, using the VTLN model trained with adults' or any other children's speech is a good alternative solution. This is in line with our previous findings [67], which indicated that VTLN models trained on Dutch improved the performance of Mandarin Chinese children's speech recognition, demonstrating the generalizability of the VTLN warp factors across languages. Overall, the best results were obtained when the VTLN model was trained on training data that consisted of speech from all three languages, all ages, and all speech types. This shows that the more variable the training data are, the better the VTLN warping factors are estimated, resulting in improved recognition performance of children's speech.

The impact of VTLN varied depending on where VTLN was applied in the automatic speech recognition process. We explored its effects when applied during both training and testing and only during testing. The approach of using VTLN during training and testing can only be applied when the model can be retrained, which is not always the case. The results show that applying VTLN only during testing gave improvements for all languages over the baseline results. Thus, even when a model cannot be retrained, applying VTLN will help children's speech recognition performance. For Dutch, applying VTLN both during training and testing gave the best results, while for German and Mandarin this condition gave slightly worse results than the test-only condition. The difference between the languages is that for Dutch, the VTLN model was trained on adults' speech with a wide variety of speech styles (including read speech, lecture recordings, broadcast data, and spontaneous conversations), while for German and Mandarin only read adults' speech was used. The results of Experiment 3 showed that the VTLN model trained on a variety of languages, speech styles, and age groups outperformed the VTLN models that were trained with less diverse data. Likewise, we hypothesize that the more diverse adult Dutch training data for the VTLN model training yielded better warping factors than the less diverse adults' speech data for German and Mandarin. This led to better normalized features, which could be learned during training, while these same normalized features were available during testing, leading to a matched train-test scenario and improved recognition performance. Importantly, the performance for adults' speech does not degrade when VTLN is applied.

The age and gender analyses on Dutch children's and teenagers' speech showed that WERs are higher for younger children and then become gradually constant with age, as shown in earlier studies that use hybrid ASR systems [6]. Although the use of VTLN maintained this trend, it improved recognition performance for younger children for all ages more compared to that of teenagers. For Dutch, the female speech was consistently recognized better than the male speech, in line with previous findings for this database [27]. The application of VTLN gave very similar improvements for both genders in the database.

Both speed perturbations and spectral augmentation are often used as data augmentation techniques in E2E and have shown their effectiveness in improving recognition performance for adults' speech [21,22] despite the fact that both methods can potentially lead to artifacts in the generated speech signal and acoustic features, respectively. Speed perturbation, for instance, alters the speech signal's pitch and speed, which occasionally

leads to unnatural or distorted sounds (as shown by a different experiments in our lab). Spectral augmentation modifies spectral characteristics; however, we do not know which spectral information is modified; thus, the model is possibly also learning artificial patterns. In this work, we did not check for these artifacts nor did we try to optimize the parameter settings of these two methods; we used the standard settings. The results shown in this paper, however, indicate that the benefit of applying SP and SpecAug is larger than the negative effect of potential artifacts. Future research could investigate whether further performance benefits could be obtained when the parameter settings are tuned to the task at hand and artifacts are removed. Regarding VTLN: While our study highlighted VTLN's impact across different languages, its applicability and integration in E2E models may encounter challenges. For instance, because VTLN needs to be trained independently and then used as a processing step after feature extraction to warp the features for training the ASR network architecture, it may not be compatible with architectures that utilize raw waveform data rather than features. As a result, integrating VTLN into such architectures requires further exploration. In the future, we intend to explore the performance of existing pre-trained models, such as Whisper, in these languages as an alternative to the baseline model without augmentations or as an alternative to the model trained using data augmentations. By doing so, we aim to investigate whether VTLN still offers additional complementary information when employed with pre-trained models that are already trained on a diverse type and even diverse speaker groups. While retraining these pre-trained models is not always feasible or desirable for computational reasons, using VTLN only during testing could potentially enhance the recognition performance of pre-trained models without extensive retraining, with the ultimate aim to remove bias against children's speech in automatic speech recognition.

In conclusion, this research contributes to narrowing the performance gap between children's and adults' speech recognition, especially when children's speech and text data are absent for training. By training our VTLN model on adults' speech and using state-of-the-art speed perturbations and spectral augmentation techniques applied to adults' speech, we improved recognition performance across diverse child speaker groups, speaking styles, and languages, thus showing that these approaches generalize across age, speaking styles, and languages. Performance was further improved when children's speech and/or highly variable speech was used to train the VTLN model. These findings highlight the potential for enhancing the End-to-End children's speech recognition performance by (1) applying state-of-the-art techniques that have shown their effectiveness on adults' speech ASR (the data augmentation techniques) and in hybrid ASR models (VTLN) to adults' speech, and (2) strategically taking into account the availability of data and the feasibility of training methods to improve children's speech recognition results in the absence of children's speech and text data for training ASR models. This finding allows for the development of more accessible and inclusive children's speech technology applications.

Author Contributions: Conceptualization, T.P.; methodology, T.P. and O.S.; software, T.P.; validation, T.P. and O.E; formal analysis, T.P.; investigation, T.P.; resources, O.S.; data curation, T.P.; writing—original draft preparation, T.P.; writing—review and editing, O.S.; visualization, T.P.; supervision, O.S. All authors have read and agreed to the published version of the manuscript.

Funding: The APC was funded by TU Delft, The Netherlands.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available upon request from the corresponding authors of the database. All databases are not publicly accessible accessible as they might be intended solely for research purposes.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AM	Acoustic Model
ASR	Automatic Speech Recognition
CER	Character Error Rate
CGN	Corpus Gesproken Nederlands
CSR	Children's Speech Recognition
CV	Common Voice
DC	Dutch Native Children
DT	Dutch Native Teenagers
E2E	End-to-End
HMI	Human–Machine Interaction
LM	Language Model
NnT	Dutch Non-Native Teenagers
SLT	Speech and Language Technology
SP	Speed Perturbations
SpecAug	Spectral Augmentation
VTLN	Vocal Tract Length Normalization
WER	Word Error Rate

Appendix A. Results of Open-AI-Whisper on Dutch, German, and Mandarin Adults' and Children's Speech

As seen in Table A1, we compared one of our best models (see, Table A1 last row) with Whisper small, medium, and large on the adults' and children's speech for all three languages. In short, for adults' speech, our model outperformed Whisper small, medium, and large despite not using a Language Model, for all three languages, except for German, where Whisper large outperformed our model, and Mandarin where Whisper medium outperformed our model. This shows that for adults' speech, our models are state-of-the-art. For children's speech, the picture is different: almost all Whisper models outperformed our model. However, it is unclear which speech data from which speaker groups were used to train Whisper, which may thus well contain children's speech. Since in this work, the aim is to investigate CSR when exclusively relying on adults' speech for training, we did not use pre-trained models for our experiments. Training our own ASR models using known data additionally offers greater flexibility in architecture and parameter tuning, allowing for a more justifiable interpretation of the results

Table A1. Performance of the Open AI-Whisper small, medium, and large models on the Dutch, German, and Mandarin adults' and children's speech test sets used in this study. For reference, row SP + SpecAug shows the (average) results of the SP + SpecAug models for the respective languages.

	Dutch (WER)				**German (WER)**		**Mandarin (CER)**		
	CGN-Read	**CGN-CTS**	**Jas-Read**	**Jas-HMI**	**CV**	**KidsTalc**	**SetA**	**SetC1**	**SetC2**
Whisper-small	17.1	54.1	39.87	51.47	9.3	58.4	12.02	7.25	13.37
Whisper-medium	12.4	39.1	30.1	41.87	5.7	40.9	9.45	4.92	11.95
Whisper-large	10.1	40.6	28.57	39.67	4.4	49.1	10.32	5.57	11.00
SP + SpecAug	7	20.2	35.3	40.2	5.1	67.2	9.9	10.0	38.8

Appendix B. Results of the Statistical Tests

Table A2. *p*-values of the performance difference between the various models in Experiments 1 and 2 and the baseline model for the Dutch adults' and children's speech.

		CGN		Jasmin: Read			Jasmin: HMI		
Augmentation	**Normalization**	**Rd**	**CTS**	**DC**	**DT**	**NnT**	**DC**	**DT**	**NnT**
SP	None	<0.001	<0.001	<0.001	<0.001	0.089	<0.001	<0.001	0.020
SP + SpecAug	None	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
None	$VTLN_{CGN}$	0.395	0.395	<0.001	<0.001	0.08	<0.001	<0.001	0.082
None	$VTLN_{Jas\text{-}DCDT}$	0.631	0.342	<0.001	0.28	0.818	<0.001	<0.001	0.019
None	$VTLN_{Jas\text{-}DCDTNnT}$	0.741	0.332	<0.001	0.001	0.001	<0.001	<0.001	<0.001

Table A3. *p*-values of the performance difference between the various models in Experiments 1 and 2 and the baseline model for the German adults' and children's speech.

Augmentation	**Normalization**	**CV**	**kidsTALC**
SP	None	<0.001	<0.001
SP + SpecAug	None	<0.001	<0.001
None	$VTLN_{CV}$	0.332	0.002
None	$VTLN_{kidsTALC}$	<0.001	<0.001

Table A4. *p*-values of the performance difference between the various models in Experiments 1 and 2 and the baseline model for the Mandarin adults' and children's speech.

Augmentation	**Normalization**	**SetA**	**SetC1**	**SetC2**
SP	None	1.000	1.000	1.000
SP + SpecAug	None	<0.001	<0.001	<0.001
None	$VTLN_{SetA}$	<0.001	<0.001	<0.001
None	$VTLN_{SetC1C2}$	<0.001	<0.001	<0.001

Table A5. *p*-values of the performance difference between the various models in Experiment 3 and the SP + SpecAug model for the Dutch adults' and children's speech.

		CGN		Jasmin: Read			Jasmin: HMI		
VTLN	**VTLNmodel**	**test_stu**	**test_tel**	**DC**	**DT**	**NnT**	**DC**	**DT**	**NnT**
Train | Test	$VTLN_{CGN}$	0.271	0.81	<0.001	<0.001	<0.001	<0.001	0.61	0.168
	$VTLN_{Jas\text{-}DCDT}$	0.006	0.007	<0.001	<0.001	0.003	<0.001	0.897	0.49
	$VTLN_{Jas\text{-}DCDTNnT}$	0.472	0.453	<0.001	<0.001	<0.001	<0.001	0.011	0.107
	$VTLN_{MultiV1}$	0.15	0.091	<0.001	<0.001	<0.001	<0.001	0.889	0.063
Test	$VTLN_{CGN}$	0.697	0.194	<0.001	0.012	<0.001	<0.001	0.002	0.294
	$VTLN_{Jas\text{-}DCDT}$	<0.001	<0.001	<0.001	<0.001	<0.001	0.002	0.006	<0.001
	$VTLN_{Jas\text{-}DCDTNnT}$	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.004	<0.001
	$VTLN_{MultiV1}$	<0.001	0.015	<0.001	0.018	<0.001	<0.001	0.121	0.064

Table A6. *p*-values of the performance difference between the various models in Experiment 3 and the SP + SpecAug model for the German adults' and children's speech.

VTLN	**VTLNmodel**	**CV**	**KidsTALC**
Train | Test	$VTLN_{CV}$	0.39	0.126
	$VTLN_{kidsTALC}$	<0.001	<0.001
	$VTLN_{MultiV1}$	0.003	0.01
Test	$VTLN_{CV}$	<0.001	<0.001
	$VTLN_{kidsTALC}$	<0.001	<0.001
	$VTLN_{MultiV1}$	<0.001	<0.001

Table A7. *p*-values of the performance difference between the various models in Experiment 3 and the SP + SpecAug model for the Mandarin adults' and children's speech.

VTLN	VTLNmodel	SetA	SetC1	SetC2
Train | Test	$VTLN_{SetA}$	0.912	0.28	0.001
	$VTLN_{SetC1C2}$	0.841	0.139	0.002
	$VTLN_{MultiV1}$	0.529	0.509	<0.001
Test	$VTLN_{SetA}$	<0.001	0.711	<0.001
	$VTLN_{SetC1C2}$	<0.001	0.061	<0.001
	$VTLN_{MultiV1}$	<0.001	0.952	<0.001

References

1. Narayanan, S.; Potamianos, A. Creating conversational interfaces for children. *IEEE Trans. Speech Audio Process.* **2002**, *10*, 65–78. [CrossRef]
2. SoapBox Labs: Speech Technology for Kids. Available online: https://www.soapboxlabs.com/ (accessed on 22 November 2023).
3. Potamianos, A.; Narayanan, S.; Lee, S. Automatic speech recognition for children. In Proceedings of the Eurospeech, Rhodes, Greece, 22–25 September 1997; pp. 2371–2374.
4. Potamianos, A.; Narayanan, S. Robust recognition of children's speech. *IEEE Trans. Speech Audio Process.* **2003**, *11*, 603–616. [CrossRef]
5. Qian, M. Computer Analysis of Children's Non-Native English Speech for Language Learning and Assessment. Ph.D. Thesis, University of Birmingham, Birmingham, UK, 2021.
6. Lee, S.; Potamianos, A.; Narayanan, S. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *J. Acous. Soc. Am.* **1999**, *105*, 1455–1468. [CrossRef] [PubMed]
7. Goldstein, U.G. Modeling children's vocal tracts. *J. Acous. Soc. Am.* **1979**, *65*, S25. [CrossRef]
8. Mermelstein, P. Articulatory model for the study of speech production. *J. Acous. Soc. Am.* **1973**, *53*, 1070–1082. [CrossRef] [PubMed]
9. Ritvo, D.; Bavitz, C.; Gupta, R.; Oberman, I. *Privacy and Children's Data—An Overview of the Children's Online Privacy Protection Act and the Family Educational Rights and Privacy Act*; Berkman Center Research Publication: Cambridge, MA, USA, 2013.
10. Voigt, P.; Von dem Bussche, A. *The EU General Data Protection Regulation (GDPR): A Practical Guide*, 1st ed.; Springer International Publishing: Cham, Switzerland, 2017.
11. Shivakumar, P.G.; Narayanan, S. End-to-End neural systems for automatic children speech recognition: An empirical study. *Comput. Speech Lang.* **2022**, *72*, 1–24.
12. Kim, D.; Umesh, M.G.S.; Hain, T.; Woodland, P. Using VTLN for broadcast news transcription. In Proceedings of the International Conference on Spoken Language Processing (ICSLP), Jeju Island, Republic of Korea, 4–8 October 2004; pp. 1953–1956.
13. Shivakumar, P.G.; Potamianos, A.; Lee, S.; Narayanan, S. Improving speech recognition for children using acoustic adaptation and pronunciation modeling. In Proceedings of the Workshop on Child, Computer and Interaction (WOCCI), Singapore, 19 September 2014.
14. Ghai, S.; Sinha, R. Exploring the role of spectral smoothing in context of children's speech recognition. In Proceedings of the Interspeech, Brighton, UK, 6–10 September 2009; pp. 1607–1610.
15. Kathania, H.K.; Shahnawazuddin, S.; Adiga, N.; Ahmad, W. Role of prosodic features on Children's Speech Recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Process, (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5519–5523.
16. Singh, V.P.; Sailor, H.; Bhattacharya, S.; Pandey, A. Spectral modification based data augmentation For improving End-to-End ASR For children's speech. In Proceedings of the Interspeech, Incheon, Republic of Korea, 18–22 September 2022; pp. 3213–3217.
17. Shahnawazuddin, S.; Adiga, N.; Kumar, K.; Poddar, A.; Ahmad, W. Voice conversion based data augmentation to improve children's speech recognition in limited data scenario. In Proceedings of the Interspeech, Shanghai, China, 25–29 October 2020; pp. 4382–4386.
18. Chen, G.; Na, X.; Wang, Y.; Yan, Z.; Zhang, J.; Ma, S.; Wang, Y. Data augmentation for Children's Speech Recognition - The "Ethiopian" system for the SLT 2021 children speech recognition challenge. *arXiv* **2020**, arXiv:abs/2011.04547.
19. Watanabe, S.; Hori, T.; Karita, S.; Hayashi, T.; Nishitoba, J.; Unno, Y.; Enrique Yalta Soplin, N.; Heymann, J.; Wiesner, M.; Chen, N.; et al. ESPnet: End-to-End Speech processing toolkit. In Proceedings of the Interspeech, Hydrebad, India, 2–6 September 2018; pp. 2207–2211.
20. Karita, S.; Chen, N.; Hayashi, T.; Hori, T.; Inaguma, H.; Jiang, Z.; Someki, M.; Soplin, N.E.Y.; Yamamoto, R.; Wang, X.; et al. A comparative study on transformer vs RNN in speech applications. In Proceedings of the Workshop on Automatic Speech Recognition and Understanding (ASRU), Sentosa, Singapore, 14–18 December 2019; pp. 449–456.
21. Ko, T.; Peddinti, V.; Povey, D.; Khudanpur, S. Audio augmentation for speech recognition. In Proceedings of the Interspeech, Dresden, Germany, 6–10 September 2015; pp. 3586–3589.

22. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. SpecAugment: A simple data augmentation method for automatic speech recognition. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 2613–2617.
23. Zhang, Y.; Herygers, A.; Patel, T.B.; Yue, Z.; Scharenborg, O. Exploring data augmentation in bias mitigation against non-native-accented speech. In Proceedings of the Workshop on Automatic Speech Recognition and Understanding (ASRU), Taipei, Taiwan, 16–20 December 2023; pp. 1–8.
24. Miao, Y.; Gowayyed, M.; Na, X.; Ko, T.; Metze, F.; Waibel, A. An empirical exploration of CTC acoustic models. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Process, (ICASSP), Shanghai, China, 20–25 March 2016; pp. 2623–2627.
25. Giuliani, D.; Gerosa, M. Investigating recognition of children's speech. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Process, (ICASSP), Shanghai, China, 20–25 March 2003; Volume 2, pp. 11–137.
26. Lansdown, G.; Vaghri, Z. Article 1: Definition of a child. In *Monitoring State Compliance with the UN Convention on the Rights of the Child: An Analysis of Attributes*; Vaghri, Z., Zermatten, J., Lansdown, G., Ruggiero, R., Eds.; Springer International Publishing: Cham, Switerland, 2022.
27. Feng, S.; Kudina, O.; Halpern, B.M.; Scharenborg, O. Quantifying bias in automatic speech recognition. *arXiv* **2021**, arXiv:2103.15122.
28. Wikipedia. Child Development. Available online: https://en.wikipedia.org/wiki/Child_development (accessed on 24 February 2024).
29. Eskenazi, M.; Mostow, J.; Graff, D. *The CMU Kids Corpus LDC97S63*; Linguistic Data Consortium: Philadelphia, PA, USA, 1997.
30. Shobaki, K.; Hosom, J.P.; Cole, R.A. The OGI kids² speech corpus and recognizers. In Proceedings of the Interspeech, Beijing, China, 16–20 October 2000; pp. 258–261.
31. Cole, R.; Hosom, P.; Pellom, B. University of colorado prompted and read children's speech corpus. In *Technical Report TR-CSLR-2006-02*; University of Colorado: Boulder, CO, USA, 2006.
32. Cole, R.; Pellom, B. University of Colorado read and summarized story corpus. In *Technical Report TR-CSLR-2006-03*; University of Colorado: Boulder, CO, USA, 2006.
33. Zhao, S.; Singh, M.; Woubie, A.; Karhila, R. Data augmentation for children ASR and child-adult speaker classification using voice conversion methods. In Proceedings of the Interspeech, Centre Dublin, Ireland, 20–24 August 2023; pp. 4593–4597.
34. Batliner, A.; Blomberg, M.; D'Arcy, S.; Elenius, D.; Giuliani, D.; Gerosa, M.; Hacker, C.; Russell, M.; Steidl, S.; Wong, M. The PFSTAR children's speech corpus. In Proceedings of the Interspeech, Lisbon, Portugal, 4–8 September 2005; pp. 2761–2764.
35. Cosi, P.; Pellom, B.L. Italian Children's Speech Recognition for advanced interactive literacy tutors. In Proceedings of the Interspeech, Lisbon, Portugal, 4–8 September 2005; pp. 2201–2204.
36. Cucchiarini, C.; Hamme, H.V.; van Herwijnen, O.; Smits, F. 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 Language Resources and Evaluation Conference (LREC), Genoa, Italy, 24–26 May 2006.
37. Yu, F.; Yao, Z.; Wang, X.; An, K.; Xie, L.; Ou, Z.; Liu, B.; Li, X.; Miao, G. The SLT 2021 children speech recognition challenge: Open datasets, rules and baselines. In Proceedings of the IEEE Speech and Language Technology (SLT), Shenzhen, China, 19–22 January 2021.
38. Rumberg, L.; Gebauer, C.; Ehlert, H.; Wallbaum, M.; Bornholt, L.; Ostermann, J.; Lüdtke, U. kidsTALC: A Corpus of 3- to 11-year-old German children's connected natural speech. In Proceedings of the Interspeech, Incheon, Republic of Korea, 18–22 September 2022; pp. 5160–5164.
39. Miller, J.; Lee, S.; Uchanski, R.; Heidbreder, A.; Richman, B.; Tadlock, J. Creation of two children's speech databases. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Process, (ICASSP), Atlanta, GA, USA, 7–10 May 1996; pp. 849–852.
40. Ward, W.; Cole, R.; Bolanos, D.B.; Buchenroth-Martin, C.; Svirsky, E.; Vuuren, S.V.; Weston, T.; Zheng, J.; Becker, L. My Science Tutor: A conversational multimedia virtual tutor for elementary school Science. *ACM Trans. Speech Lang. Process.* **2011**, *7*, 1–29. [CrossRef]
41. Gretter, R.; Matassoni, M.; Bannò, S.; Falavigna, D. TLT-school: A corpus of non native children speech. In Proceedings of the Language Resources and Evaluation Conference LREC, Marseille, France, 11–16 May 2020; pp. 378–385.
42. Interspeech. Special Session: Connecting Speech Science and Speech Technology for Children's Speech. 2023. Available online: https://sites.google.com/view/sciencetech4childspeech-is23 (accessed on 24 February 2024).
43. Interspeech. Special Session: Shared Task on Automatic Speech Recognition for Non-Native Children's Speech. 2020. Available online: https://sites.google.com/view/wocci/home/interspeech-2020-special-session (accessed on 24 February 2024).
44. Interspeech. Special Session: Shared Task on Automatic Speech Recognition for Non-Native Children's Speech. 2021. Available online: https://sites.google.com/fbk.eu/ss-is2021-nonnativechildren/ (accessed on 24 February 2024).
45. Potamianos, A.; Narayanan, S. Spoken dialog systems for children. In Proceedings of the IEEE International Conference. on Acoustics, Speech and Signal Process, (ICASSP), Seattle, WA, USA, 12–15 May 1998; pp. 197–200.
46. Kazemzadeh, A.; You, H.; Iseli, M.; Jones, B.; Cui, X.; Heritage, M.; Price, P.; Anderson, E.; Narayanan, S.; Alwan, A. TBALL data collection: The making of a young children's speech corpus. In Proceedings of the Interspeech, Lisbon, Portugal, 4–8 September 2005; pp. 1581–1584.

47. Shahnawazuddin, S.; Dey, A.; Sinha, R. Pitch-Adaptive front-end features for robust children's ASR. In Proceedings of the Interspeech, San Francisco, CA, USA, 8–12 September 2016; pp. 3459–3463.
48. Sivaraman, G.; Mitra, V.; Nam, H.; Tiede, M.; Espy-Wilson, C. Vocal tract length normalization for speaker independent acoustic-to-articulatory speech inversion. In Proceedings of the Interspeech, San Francisco, CA, USA, 8–12 September 2016; pp. 455–459.
49. Cosi, P. On the development of matched and mismatched Italian children's speech recognition systems. In Proceedings of the Interspeech, Brighton, UK, 6–10 September 2009; pp. 540–543.
50. Gurunath Shivakumar, P.; Georgiou, P. Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations. *Comput. Speech Lang. (CSL)* **2020**, *63*, 101077. [CrossRef] [PubMed]
51. Kathania, H.K.; Kadiri, S.R.; Alku, P.; Kurimo, M. A formant modification method for improved ASR of children's speech. *Speech Commun.* **2022**, *136*, 98–106. [CrossRef]
52. Ng, S.I.; Liu, W.; Peng, Z.; Feng, S.; Huang, H.P.; Scharenborg, O.; Lee, T. The CUHK-TUDelft system for the SLT 2021 children speech recognition challenge. *arXiv* **2020**, arXiv:2011.06239.
53. Gelin, L.; Pellegrini, T.; Pinquier, J.; Daniel, M. Simulating reading mistakes for child speech transformer-based phone recognition. In Proceedings of the Interspeech, Brno, Czech Republic, 30 August–3 September 2021; pp. 3860–3864.
54. Sinha, R.; Ghai, S. On the use of pitch normalization for improving children's speech recognition. In Proceedings of the Interspeech, Brighton, UK, 6–10 September 2009; pp. 568–571.
55. Jain, R.; Barcovschi, A.; Yiwere, M.; Corcoran, P.; Cucu, H. Adaptation of Whisper models to child speech recognition. In Proceedings of the Interspeech, Dublin, Ireland, 20–24 August 2023; pp. 5242–5246.
56. Oostdijk, N. The spoken Dutch corpus. Overview and first evaluation. In Proceedings of the Language Resources and Evaluation Conference (LREC), Athens, Greece, 31 May–2 June 2000; pp. 887–894.
57. van Leeuwen, D.A.; Kessens, J.; Sanders, E.; van den Heuvel, H. Results of the n-best 2008 dutch speech recognition evaluation. In Proceedings of the Interspeech, Brighton, UK, 6–10 September 2009; pp. 2571–2574.
58. Ardila, R.; Branson, M.; Davis, K.; Henretty, M.; Kohler, M.; Meyer, J.; Morais, R.; Saunders, L.; Tyers, F.M.; Weber, G. Common Voice: A massively-multilingual speech corpus. In Proceedings of the Language Resources and Evaluation Conference (LREC), Marseille, France, 11–16 May 2020; pp. 4211–4215.
59. Li, Z. Mitigating Regional Accent Bias in ASR Systems. Master's Thesis, TU Delft, Delft, The Netherlands, 2023.
60. Sarkar, A.K.; Rath, S.P.; Umesh, S. Vocal tract length normalization factor based speaker-cluster UBM for speaker verification. In Proceedings of the National Conference on Communications (NCC), Chennai, India, 29–31 January 2010; pp. 1–5.
61. Gales, M.; Young, S. The application of hidden Markov models in speech recognition. *Found. Trends Signal Process.* **2007**, *1*, 195–304. [CrossRef]
62. Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M.; Motlicek, P.; Qian, Y.; Schwarz, P.; et al. The Kaldi speech recognition toolkit. In Proceedings of the Workshop on Automatic Speech Recognition and Understanding (ASRU), Waikoloa, HI, USA, 11–15 December 2011.
63. Gulati, A.; Qin, J.; Chiu, C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; et al. Conformer: Convolution-augmented transformer for speech recognition. In Proceedings of the Interspeech, Shanghai, China, 25–29 October 2020; pp. 5036–5040.
64. Guo, P.; Boyer, F.; Chang, X.; Hayashi, T.; Higuchi, Y.; Inaguma, H.; Kamo, N.; Li, C.; Garcia-Romero, D.; Shi, J.; et al. Recent developments on ESPnet toolkit boosted by conformer. *arXiv* **2020**, arXiv:2010.13956.
65. The ASR Training Recipes and Scripts. 2024. Available online: https://github.com/tanvinabpatel/E2ECSR-Methods/ (accessed on 24 February 2024).
66. Radford, A.; Kim, J.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust speech recognition via large-scale weak supervision. In Proceedings of the International Conference on Machine Learning (ICML), Honolulu, HI, USA, 23–29 July 2023; pp. 28492–28518.
67. Patel, T.; Scharenborg, O. Using data augmentations and VTLN to reduce bias in Dutch End-to-End speech recognition systems. *arXiv* **2023**, arXiv:2307.02009.
68. Park, Y.; Patwardhan, S.; Visweswariah, K.; Gates, S.C. An empirical analysis of word error rate and keyword error rate. In Proceedings of the Interspeech, Brisbane, Australia, 22–26 September 2008; pp. 2070–2073.
69. WER Statistical Significance Test. Available online: https://github.com/talhanai/wer-sigtest (accessed on 24 February 2024).
70. Pallet, D.; Fisher, W.; Fiscus, J. Tools for the analysis of benchmark speech recognition tests. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Process, (ICASSP), Albuquerque, NM, USA, 3–6 April 1990.
71. Feng, S.; Halpern, B.M.; Kudina, O.; Scharenborg, O. Towards inclusive automatic speech recognition. *Compututer Speech Language (CSL)* **2024**, *84*, 101567. [CrossRef]
72. Gebauer, C.; Rumberg, L.; Ehlert, H.; Lüdtke, U.; Ostermann, J. Exploiting diversity of automatic transcripts from distinct speech recognition techniques for children's speech. In Proceedings of the Interspeech, Dublin, Ireland, 20–24 August 2023; pp. 4578–4582.
73. Ghai, S.; Sinha, R. *Adaptive Feature Truncation to Address Acoustic Mismatch in Automatic Recognition of Children's Speech*; Asia-Pacific Signal and Information Processing (APSIPA): Jeju, Republic of Korea, 2016; Volume 5, p. e15.
74. Long, M.H. Maturational constraints on language development. *Stud. Second. Lang. Acquis.* **1990**, *12*, 251–285. [CrossRef]

75. MAGICDATA Mandarin Chinese Read Speech Corpus. Available online: https://www.openslr.org/68/ (accessed on 24 February 2024).
76. Bu, H.; Du, J.; Na, X.; Wu, B.; Zheng, H. AIShell-1: An Open-Source Mandarin speech corpus and a speech recognition baseline. In Proceedings of the Oriental COCOSDA, Seoul, Republic of Korea, 1–3 November 2017.
77. Wang, D.; Zhang, X.; Zhang, Z. THCHS-30: A free Chinese speech corpus. *arXiv* **2015**, arXiv:1512.01882.
78. Lu, R.; Shahin, M.; Ahmed, B. Improving Children's Speech Recognition by fine-tuning self-supervised adults' speech representations. *arXiv* **2022**, arXiv:2211.07769.