Automatic Speech Recognition for Children: A Systematic Review of Models, Toolkits, and Adaptations

Raul Benites Paradeda[1], **Karine Bezerra Furtado**[1], **Giulia de Oliveira Moscoso**[1]

[1]Department of Computer Science
State University of Rio Grande do Norte – Natal/RN– Brazil

raulparadeda@uern.br,

karinefurtado,giuliamoscoso@alu.uern.br

Abstract. *Automatic Speech Recognition (ASR) has emerged as a transformative technology in education, yet its application to children's speech remains underexplored. This paper presents a systematic review of ASR for children, focusing on its effectiveness in educational contexts. Following the guidelines of Kitchenham and Charters (2007), we analyzed academic articles published between 2019 and 2024, sourced from the ACM Digital Library, IEEE Xplore, and Scopus. A total of 16 articles were selected based on predefined inclusion criteria, including relevance to children's speech and educational applications. The review identifies Deep Neural Networks (DNNs) and adversarial learning models as the most effective approaches for recognizing children's speech. Key findings highlight the potential of ASR to enhance language learning and development in children, particularly in low-resource contexts. However, challenges such as data scarcity and the need for adaptation to diverse linguistic environments remain significant barriers. This study contributes to the ongoing discussion on innovative educational technologies by providing a comprehensive analysis of current trends and future directions in ASR for children.*

Keywords: Automatic Speech Recognition, Children's Speech, Educational Technology, Language Development, Systematic Review.

1. Introduction

Automatic Speech Recognition (ASR), defined as the process of converting speech signals into text [Li et al. 2015], has emerged as a transformative technology in education. By enabling seamless interaction between humans and machines, ASR has the potential to create more inclusive, accessible, and engaging learning environments. While significant progress has been made in ASR for adults, the application of this technology to children's speech remains underexplored, presenting a critical gap in the development of educational tools tailored to younger learners.

The COVID-19 pandemic accelerated the adoption of digital learning platforms, highlighting the importance of technologies like ASR in education. From virtual tutoring to language learning apps, ASR has proven essential in enhancing accessibility, particularly for students with hearing impairments [Aljedaani et al. 2023]. For children with learning difficulties such as dyslexia or challenges in handwriting and spelling, ASR offers a transformative solution by enabling voice-to-text functionality, thereby reducing

barriers to written expression [Husni and Jamaludin 2009]. As children increasingly engage with digital devices for education, gaming, and social interaction [UNICEF 2017], the development of ASR systems adapted to their unique needs is crucial for fostering effective human-machine interaction in educational contexts.

This is particularly relevant in computing education, where initiatives to incorporate computational thinking into basic education curricula are gaining momentum in Brazil and globally. ASR technologies can provide more accessible pathways to programming concepts by allowing children to engage with computational tools through natural speech, supporting their first interactions with coding before they fully develop traditional typing skills or syntax comprehension.

However, recognizing children's speech poses distinct challenges. Variations in vocal tract morphology, limited control over prosodic elements like pitch and intonation, and rapid developmental changes result in lower performance when conventional ASR systems are applied to children's speech [Bhardwaj et al. 2022, Southwell et al. 2022]. These challenges underscore the need for specialized approaches in ASR technology to ensure its effectiveness in educational settings for children.

To address these challenges and explore the potential of ASR in education, this paper presents a systematic review of ASR for children, focusing on its applications in educational contexts. By analyzing research articles published between 2019 and 2024, we aim to identify trends, evaluate the effectiveness of current ASR models, and explore how these technologies are being integrated into learning environments.

This systematic literature review synthesizes existing knowledge and identifies key research focus areas at the intersection of ASR, child education, and educational technology. By comprehensively examining the literature, we seek to clarify the diversity of acoustic models used, the tools employed in the recognition process, and how these technologies are being integrated into educational environments to enhance learning and inclusion. By addressing these questions, this study aims to inform future research and development of ASR systems that are effective, inclusive, and tailored to the needs of young learners.

2. Background

Understanding the foundational technologies and challenges of ASR is crucial, especially for children's speech, which presents unique phonetic and linguistic traits. This section outlines the key components of speech technology and emphasizes the need to adapt ASR systems to children's specific characteristics in educational contexts.

2.1. Speech Technology in Education

Speech Technology, an interdisciplinary field encompassing electrical engineering, signal processing, computational linguistics, and language psychology, has become increasingly relevant in educational contexts. Its primary objective is to develop systems capable of understanding and processing human speech, transforming it into digital information useful for various educational applications [Li et al. 2022].

ASR, a key application of speech technology, enables more intuitive interaction with digital platforms. It allows students to provide voice input, enhancing pronunciation

skills and language learning through real-time feedback. Research has shown that ASR can significantly improve segmental pronunciation (sounds within words), with a moderate impact on suprasegmental features like intonation and rhythm, making it an essential tool for language learners [Ngo et al. 2024]. Furthermore, ASR applications are valuable in making learning more accessible, particularly for students with physical or cognitive disabilities who may struggle with traditional text-based interfaces [Ngo et al. 2024].

2.2. Automatic Speech Recognition (ASR)

Automatic Speech Recognition, also known as speech-to-text (STT) or computer speech recognition, is a key component of speech technology [Gold et al. 2011]. ASR systems process spoken words from audio files, converting them into text for computers to interpret natural language. In educational settings, ASR has numerous applications, including language learning, reading assistance, accessibility for students with hearing impairments, and interactive learning environments [Alyoussef 2021].

ASR systems are trained on large datasets to create models for speech analysis, which can be tailored to different speech types, speakers, vocabularies, and environmental factors. This adaptability is crucial in educational contexts, where the system may need to handle diverse accents, age groups, and learning environments [Li et al. 2015].

2.3. Challenges in Children's Speech Recognition

For applications aimed at children, ASR presents unique challenges due to the nature of children's speech [Nagano et al. 2019]. These challenges include variations in pronunciation, intonation, speech rate, vocabulary, and grammatical structure compared to adult speech [Potamianos and Narayanan 2003]. In educational technology, these differences necessitate the development of specialized ASR systems that can effectively recognize and process children's speech patterns.

Key challenges in developing ASR for educational tools targeting children:

- Higher variability in acoustic properties of speech;
- Limited vocabulary and simpler grammatical structures;
- Unpredictable speech patterns and pronunciation errors;
- Rapid developmental changes affecting speech characteristics.

Addressing these challenges is crucial for creating effective educational technologies that can accurately interpret and respond to children's speech input.

2.4. Resources for Automatic Speech Recognition

A speech collection, also known as a corpus, comprises audio recordings of spoken language paired with written transcriptions. Within the speech recognition domain, these corpora serve purposes such as acoustic analysis and the creation of models to identify speech patterns and speakers. The presence of ample publicly accessible speech corpora significantly influences advancements in speech recognition research. Most speech corpora designed for recognizing children's speech have been curated with input from individuals aged 5 to 18 years [Claus et al. 2013]. Among the widely utilized publicly accessible corpora for modeling acoustics in children's speech across various languages, some notable examples include:

- **PF-STAR children's speech corpora**: A corpus designed for studying children's speech in English, particularly for speech recognition and synthesis [Russell 2006].
- **TIDIGITS**: A dataset of spoken digits, widely used for testing ASR systems, including those for children [Gary 1993].
- **CMU Kids corpora**: A dataset from Carnegie Mellon University, featuring children's speech in English for ASR development [Eskenazi et al. 1997].
- **EmoChildRu**: A corpus of children's emotional speech in Russian, useful for studying prosody and emotion in ASR [Lyakso et al. 2015].

Important to note that, despite the availability of corpora in multiple languages, we did not find a dedicated corpus for Portuguese, highlighting a gap in resources for ASR research in this language.

2.4.1. ASR Toolkits

In the area of research and development, particularly in creating ASR applications, researchers have access to open-source speech recognition toolkits for crafting sophisticated systems. Among the most widely used toolkits are:

- **SpeechBrain**: A Python-based toolkit for neural network-based ASR systems.
- **Kaldi**: A C++ toolkit widely used for developing neural network-based ASR models.
- **OpenSMILE**: A C++ toolkit focused on extracting Mel-Frequency Cepstral Coefficients (MFCCs) for speech analysis.

These toolkits provide researchers with the necessary tools to develop and optimize ASR systems, catering to different programming languages and technical conditions.

2.5. Children's vs. Adults' Speech: Educational Implications

Understanding the differences between children's and adults' speech is crucial for developing effective educational technologies using ASR [Taniya et al. 2020]. As children grow, their speech patterns evolve, affecting both acoustic and linguistic aspects [Chermak and Schneiderman 1986]. Key differences include:

- **Acoustic Properties**: Children typically have higher formant[1] and fundamental frequencies due to their smaller vocal folds and shorter vocal tracts [Kathania et al. 2021].
- **Developmental Changes**: Studies have observed changes in formant and fundamental frequencies among child speakers aged three to thirteen [Coughler et al. 2022].
- **Language Skills**: Generally, children's language skills improve with age, with speech becoming more fluent around ages 12 to 13 [Liberman 1989].
- **Pronunciation**: Between ages 8 and 10, the likelihood of mispronouncing words is significantly higher compared to 11- to 14-year-olds [Assmann et al. 2013].

[1] Formant is a resonant frequency peak in the vocal tract's acoustic spectrum.

These differences in children's speech development have crucial implications for the design and application of educational technology:

- **Adaptability**: ASR systems in educational tools must be adaptable to account for the rapid developmental changes that occur in children's speech [Potamianos and Narayanan 2003].
- **Pronunciation and Variability**: Educational software incorporating ASR should be designed to handle the increased variability in children's pronunciation and speech patterns [Gerosa et al. 2007].
- **Age-Appropriate Feedback**: Feedback mechanisms in language learning applications need to consider age-appropriate expectations for pronunciation and fluency development [Neri et al. 2003].
- **Multi-Age Platforms**: Educational platforms targeting children across different age ranges may require separate ASR models or adjustable settings to accommodate the various developmental stages [Wilks et al. 2010].

By considering these factors, developers can create more effective and inclusive educational technologies that accurately recognize and respond to children's speech across different age groups and developmental stages.

3. Methodology

This systematic literature review follows the guidelines by [Kitchenham and Charters 2007] and is conducted in three phases: Planning, Conducting, and Reporting.

3.1. Planning Phase

The planning phase involved defining the research questions, developing the review protocol, and identifying the search strategy.

3.1.1. Research Questions

The research questions (RQs) guiding this review are: (1) What types of acoustic models are most effective for children's speech recognition? (2) Which toolkits and frameworks are commonly used in this domain? and (3) How can ASR be adapted to support language learning and development in children, particularly in low-resource contexts?

3.1.2. Review Protocol

A review protocol was developed, including the search strategy, inclusion/exclusion criteria, data extraction process, and analysis methods. All authors approved the protocol before the search began.

3.1.3. Search Strategy

We selected three databases for their extensive collections in speech technology and education:

- **ACM Digital Library**: A respected source for computer science research, particularly related to algorithms, AI, and human-computer interaction.
- **IEEE Xplore**: Essential for technology-related research, particularly in electrical engineering and computer science.
- **Scopus**: A large multidisciplinary database widely used for literature reviews across fields.

The search string was developed according to the following steps:

1. Identify the essential terms linked to the research questions.
2. Incorporate alternative keywords or synonyms for the main terms.
3. Use Boolean operators (AND, OR) to refine the search and narrow down the results.

Based on that, the search string formulated for this review was:

automatic speech recognition AND ASR AND acoustic model AND education AND (intitle:CHILDREN OR intitle:Child OR intitle:Kids)

3.2. Conducting Phase

The conducting phase involved executing the search, applying the inclusion/exclusion criteria, and extracting and analyzing the data.

3.2.1. Study Selection

The search process yielded a total of 238 articles: 13 from the ACM Digital Library, 184 from IEEE Xplore, and 41 from Scopus. These articles were then filtered according to the following steps:

- Date Filtering: We applied a date filter to focus on the most recent advancements, retaining only studies published between 2019 and 2024. This reduced the number of articles to 35.
- Removal of Duplicates and Review Papers: Duplicate articles and review papers were removed to ensure the inclusion of only original empirical studies. This step resulted in 4 articles from the ACM Digital Library, 9 from IEEE Xplore, and 18 from Scopus, totaling 31 articles.
- Title and Abstract Screening: The titles and abstracts of the remaining articles were screened to exclude those that did not focus on ASR for children's speech. Studies that addressed adult speech, voice conversion technologies, or unrelated applications (e.g., robot interactions, and keyword recognition) were excluded. This screening process resulted in a final selection of 16 articles for full-text review.

The selection process was conducted independently by two reviewers, with discrepancies resolved through discussion or consultation with a third reviewer. Figure 1 presents the selection phase flowchart, inspired by the PRISMA method, which is widely recognized for structuring systematic reviews, though adapted according to the authors' original protocol.

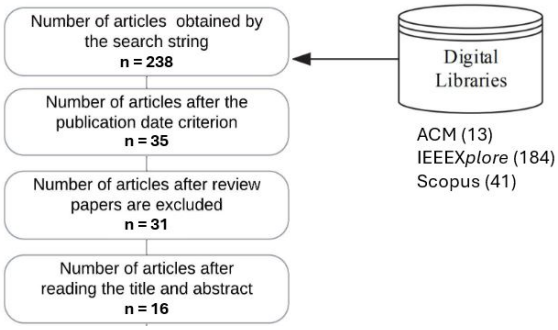

Figure 1. Adapted PRISMA flowchart.

3.2.2. Quality Assessment

The quality of the selected studies was assessed using a checklist based on relevance, methodological rigour, and clarity of findings. Studies not meeting the minimum threshold were excluded.

3.2.3. Data Extraction

Two reviewers independently extracted data on acoustic models, toolkits, and languages using a structured spreadsheet.

3.2.4. Data Synthesis

The extracted data were synthesized to identify trends, evaluate the effectiveness of current ASR models, and explore how these technologies are being integrated into educational environments. The synthesis focused on addressing the research questions and comparing the findings with those of other systematic reviews in the field.

3.3. Reporting Phase

The findings and conclusions were documented, and implications for future research and practice were discussed. Table 3.3 lists the selected studies.

4. Results

This section presents the findings for each research question. Our selection comprised one paper published in 2019, one in 2020, eight in 2021, two in 2022, and four in 2023.

4.1. RQ1: What types of models were employed for training and testing in the studies?

The analysis of the selected studies reveals a diverse range of models and techniques employed for training and testing ASR systems for children's speech. These models can be categorized into several key types:

- **Deep Neural Networks (DNNs):** Widely used for their efficiency and accuracy, as demonstrated in low-latency systems like Dorado et al.'s [Dorado and Villanueva 2023] ASR for Filipino children.

Table 1. Selected research papers

Paper	Title	Database
[Hair et al. 2019]	Evaluating Automatic Speech Recognition for Child Speech Therapy Applications	ACM
[Kothalkar et al. 2021]	Measuring Frequency of Child-directed WH-Question Words for Alternate Preschool Locations using Speech Recognition and Location Tracking Technologies	ACM
[Hair et al. 2021a]	A Longitudinal Evaluation of Tablet-Based Child Speech Therapy with Apraxia World	ACM
[Garg et al. 2022]	The Last Decade of HCI Research on Children and Voice-based Conversational Agents	ACM
[Jain et al. 2022]	A Text-to-Speech Pipeline, Evaluation Methodology, and Initial Fine-Tuning Results for Child Speech Synthesis	IEEE*Xplorer*
[Abion et al. 2023]	Comparison of Data Augmentation Techniques on Filipino ASR for Children's Speech	Scopus
[Duan 2023]	Joint Learning Feature and Model Adaptation for Unsupervised Acoustic Modelling of Child Speech	Scopus
[Getman et al. 2023]	Developing an AI-Assisted Low-Resource Spoken Language Learning App for Children	Scopus
[Dorado and Villanueva 2023]	Development of Low-Latency and Real-Time Filipino Children Automatic Speech Recognition System using Deep Neural Network	Scopus
[Lileikyte et al. 2022]	Assessing Child Communication Engagement and Statistical Speech Patterns for American English	Scopus
[Misra et al. 2021]	A Good Start is Half the Battle Won: Unsupervised Pre-training for Low Resource Children's Speech Recognition	Scopus
[Hair et al. 2021b]	Assessing Posterior-Based Mispronunciation Detection on Field-Collected Recordings from Child Speech Therapy Sessions	Scopus
[Gelin et al. 2021]	Simulating Reading Mistakes for Child Speech Transformer-Based Phone Recognition	Scopus
[Yeung et al. 2021]	Fundamental Frequency Feature Normalization and Data Augmentation for Child Speech Recognition	Scopus
[Duan and Chen 2021]	Senone-Aware Adversarial Multi-Task Training for Unsupervised Child to Adult Speech Adaptation	Scopus
[Duan and Chen 2020]	Unsupervised Feature Adaptation Using Adversarial Multi-Task Training for Automatic Evaluation of Children's Speech	Scopus

- **Transformer-based Models:** Effective in sequence tasks, such as Gelin et al.'s [Gelin et al. 2021] application for reading error simulation.
- **Adversarial/Multi-task Learning:** Duan et al. [Duan and Chen 2021] applied these techniques for unsupervised adaptation and evaluation of children's speech, addressing limited data.
- **Unsupervised Models:** Approaches like Duan et al.'s [Duan 2023] unsupervised acoustic modeling and Misra et al.'s [Misra et al. 2021] pre-training tackled data scarcity.
- **Mispronunciation Detection:** Models like Hair et al. [Hair et al. 2021b] focused on detecting pronunciation errors in child speech therapy.
- **Data Augmentation:** Studies such as Abion et al. [Abion et al. 2023] and Yeung et al. [Yeung et al. 2021] enhanced ASR accuracy through data augmentation techniques.
- **Text-to-Speech and Engagement Models:** Jain et al. [Jain et al. 2022] developed child TTS pipelines, while Lileikyte et al. [Lileikyte et al. 2022] assessed engagement in ASR interactions.
- **Low-Latency Models:** Real-time ASR systems, like those by Dorado et al. [Dorado and Villanueva 2023] and Getman et al. [Getman et al. 2023], were designed for interactive educational and therapeutic uses.

4.2. RQ2: Which Toolkits researchers use for speech recognition?

The analysis of the selected studies reveals a diverse range of toolkits and frameworks employed by researchers in the field of child speech recognition. These toolkits can be categorized as presented in Table 2.

4.3. RQ3: What are the various natural languages employed in developing the ASR system?

The analysis of the selected studies reveals that researchers have developed ASR systems for children's speech in various languages. While English appears to be the most common, there is notable effort to develop systems for other languages, particularly those considered low-resource. The languages identified in the studies are show in Table 3.

Table 2. Toolkits and Frameworks Used in Child Speech Recognition Studies

Toolkit/Framework	Usage/Study
Kaldi	- Abion et al. [Abion et al. 2023]: Data augmentation for Filipino ASR children. - Yeung et al. [Yeung et al. 2021]: Feature normalization. - Duan et al. [Duan and Chen 2021]: Adversarial multi-task training.
ESPnet	- Gelin et al. [Gelin et al. 2021]: Transformer-based phone recognition for child speech.
TensorFlow	- Dorado et al. [Dorado and Villanueva 2023]: Real-time Filipino children ASR system.
PyTorch	- Jain et al. [Jain et al. 2022]: Text-to-speech pipeline for child speech synthesis.
HTK (Hidden Markov Model)	Not explicitly mentioned but widely known in speech recognition for HMM-based models.
Mozilla DeepSpeech	Open-source ASR engine, potentially used in end-to-end systems.
Custom Frameworks	- Hair et al. [Hair et al. 2021b]: Mispronunciation detection for child speech therapy. - Getman et al. [Getman et al. 2023]: AI-assisted app for child language learning.
Commercial APIs	Services like Google Cloud Speech-to-Text and Amazon Transcribe are sometimes used for applied ASR research.

5. Discussion and Conclusions

Our analysis provides key insights into the field of ASR for children's speech, particularly regarding the types of models, toolkits, and languages employed. The dominance of deep neural networks (DNNs), transformer-based models, and adversarial/multi-task learning models in child ASR research reflects a broader trend in the field, where these approaches have demonstrated superior performance in handling complex speech patterns. This aligns with findings from previous studies (e.g., [Shahin et al. 2022, Dorado and Villanueva 2023]), which highlight the advantages of transformers in improving ASR accuracy, especially for low-resource languages and real-time applications.

One notable finding is the increasing use of adversarial and multi-task learning models to compensate for limited labeled data, a trend that has been similarly reported by [Shinohara 2016]. The success of these techniques in adapting child speech to adult ASR models demonstrates their potential for cross-domain knowledge transfer, mitigating data scarcity issues. This is particularly relevant in the context of languages with limited speech datasets, reinforcing the need for further exploration of unsupervised learning techniques in child ASR.

Another key trend is the shift towards real-time, low-latency ASR systems, as emphasized by [Dorado and Villanueva 2023]. These systems are crucial for interactive applications such as educational tools and speech therapy, where fast response times are essential. Our findings suggest that future research should prioritize optimizing ASR

Table 3. Languages Used in ASR System Development for Children's Speech

Language	Studies and Focus
English	- Hair et al. [Hair et al. 2019, Hair et al. 2021a]: Child speech therapy applications. - Lileikyte et al. [Lileikyte et al. 2022]: Assessment of American English communication engagement. - Likely used in studies by Garg et al. [Garg et al. 2022], Jain et al. [Jain et al. 2022], Hair et al. [Hair et al. 2021b].
Filipino (Tagalog)	- Abion et al. [Abion et al. 2023]: Data augmentation techniques for Filipino child ASR. - Dorado et al. [Dorado and Villanueva 2023]: Low-latency, real-time Filipino children ASR.
Low-Resource Languages	- Getman et al. [Getman et al. 2023]: AI-assisted app for low-resource spoken language learning. - Misra et al. [Misra et al. 2021]: Unsupervised pre-training for low-resource child speech recognition.
Multiple/ Unspecified	- Duan et al. [Duan 2023, Duan and Chen 2021, Duan and Chen 2020]: Unsupervised/adversarial training for various languages. - Yeung et al. [Yeung et al. 2021]: Feature normalization techniques. - Gelin et al. [Gelin et al. 2021]: Phone recognition for multiple languages.

models for real-world deployment, ensuring both efficiency and accuracy in dynamic environments.

In terms of ASR toolkits, our review found that Kaldi and ESPnet remain the most widely used frameworks, consistent with prior research in ASR system development [Watanabe et al. 2018]. However, the increased use of TensorFlow and PyTorch for implementing custom neural architectures suggests a growing trend toward greater flexibility in model design. This shift may indicate a move beyond traditional ASR toolkits toward more adaptable, end-to-end deep learning solutions.

A significant challenge identified in our review is the limited availability of ASR systems for languages other than English. While studies such as [Dorado and Villanueva 2023] and [Abion et al. 2023] have made progress in developing ASR models for Filipino, most research continues to focus on English-speaking children. This underscores a critical gap in the field: the need for more inclusive speech technologies that address the linguistic diversity of global populations. Future work should explore strategies for developing ASR models that generalize better across multiple languages, leveraging cross-lingual and multilingual training approaches.

5.1. Concluding Remarks

Our findings highlight the growing role of ASR systems in enhancing educational experiences, particularly in personalized learning and language development. The success of

deep learning approaches, including transformers and adversarial training, suggests that child ASR research is moving toward increasingly robust and adaptable models. However, challenges such as data scarcity and the computational demands of advanced models remain obstacles to widespread implementation, particularly in resource-limited educational settings.

The implications of this study extend beyond ASR technology itself, emphasizing the potential of these systems in inclusive and accessible education. Future research should focus on improving ASR performance for underrepresented languages and optimizing models for real-time, practical applications in schools and therapeutic settings. By addressing these challenges, ASR systems can contribute to more engaging, equitable, and effective learning environments worldwide.

References

[Abion et al. 2023] Abion, C., Lumapag, N. C., Ramirez, J. C., Resulto, C., and Lucas, C. R. (2023). Comparison of data augmentation techniques on filipino asr for children's speech. In *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 60–65. IEEE.

[Aljedaani et al. 2023] Aljedaani, W., Krasniqi, R., Aljedaani, S., Mkaouer, M. W., Ludi, S., and Al-Raddah, K. (2023). If online learning works for you, what about deaf students? emerging challenges of online learning for deaf and hearing-impaired students during covid-19: a literature review. *Universal access in the information society*, 22(3):1027–1046.

[Alyoussef 2021] Alyoussef, I. (2021). E-learning system use during emergency: an empirical study during the covid-19 pandemic. In *Frontiers in Education*, volume 6, page 677753. Frontiers Media SA.

[Assmann et al. 2013] Assmann, P. F., Nearey, T. M., and Bharadwaj, S. V. (2013). Developmental patterns in children's speech: Patterns of spectral change in vowels. *Vowel inherent spectral change*, pages 199–230.

[Bhardwaj et al. 2022] Bhardwaj, V., Ben Othman, M. T., Kukreja, V., Belkhier, Y., Bajaj, M., Goud, B. S., Rehman, A. U., Shafiq, M., and Hamam, H. (2022). Automatic speech recognition (asr) systems for children: A systematic literature review. *Applied Sciences*, 12(9):4419.

[Chermak and Schneiderman 1986] Chermak, G. D. and Schneiderman, C. R. (1986). Speech timing variability of children and adults. *Journal of Phonetics*, 13(4):477–80.

[Claus et al. 2013] Claus, F., Gamboa-Rosales, H., Petrick, R., Hain, H.-U., and Hoffmann, R. (2013). A survey about databases of children's speech.

[Coughler et al. 2022] Coughler, C., Quinn de Launay, K. L., Purcell, D. W., Oram Cardy, J., and Beal, D. S. (2022). Pediatric responses to fundamental and formant frequency altered auditory feedback: A scoping review. *Frontiers in Human Neuroscience*, 16:858863.

[Dorado and Villanueva 2023] Dorado, B. and Villanueva, A. (2023). Development of low-latency and real-time filipino children automatic speech recognition system using deep

neural network. *ISDFS 2023 - 11th International Symposium on Digital Forensics and Security*. Cited by: 1.

[Duan 2023] Duan, R. (2023). Joint learning feature and model adaptation for unsupervised acoustic modelling of child speech. Cited by: 0.

[Duan and Chen 2020] Duan, R. and Chen, N. F. (2020). Unsupervised feature adaptation using adversarial multi-task training for automatic evaluation of children's speech. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2020-October:3037 – 3041. Cited by: 13.

[Duan and Chen 2021] Duan, R. and Chen, N. F. (2021). Senone-aware adversarial multi-task training for unsupervised child to adult speech adaptation. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2021-June:7758 – 7762. Cited by: 7; All Open Access, Green Open Access.

[Eskenazi et al. 1997] Eskenazi, M., Mostow, J., and Graff, D. (1997). The cmu kids corpus. *Linguistic Data Consortium*, 11.

[Garg et al. 2022] Garg, R., Cui, H., Seligson, S., Zhang, B., Porcheron, M., Clark, L., Cowan, B. R., and Beneteau, E. (2022). The last decade of hci research on children and voice-based conversational agents. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

[Gary 1993] Gary, R. (1993). Leonard and george doddington. In *TIDIGITS speech corpus*. Texas Instruments, Inc.

[Gelin et al. 2021] Gelin, L., Pellegrini, T., Pinquier, J., and Daniel, M. (2021). Simulating reading mistakes for child speech transformer-based phone recognition. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 3:1918 – 1922. Cited by: 1; All Open Access, Green Open Access.

[Gerosa et al. 2007] Gerosa, M., Giuliani, D., and Brugnara, F. (2007). Analyzing children's speech: An acoustic study of consonants and consonant-vowel transition. In *INTERSPEECH*.

[Getman et al. 2023] Getman, Y., Phan, N., Al-Ghezi, R., Voskoboinik, E., Singh, M., Grosz, T., Kurimo, M., Salvi, G., Svendsen, T., Strombergsson, S., Smolander, A., and Ylinen, S. (2023). Developing an ai-assisted low-resource spoken language learning app for children. *IEEE Access*, 11:86025 – 86037. Cited by: 5; All Open Access, Gold Open Access.

[Gold et al. 2011] Gold, B., Morgan, N., and Ellis, D. (2011). *Speech and audio signal processing: processing and perception of speech and music*. John Wiley & Sons.

[Hair et al. 2019] Hair, A., Ballard, K. J., Ahmed, B., and Gutierrez-Osuna, R. (2019). Evaluating automatic speech recognition for child speech therapy applications. In *Proceedings of the 21st International ACM SIGACCESS conference on computers and accessibility*, pages 578–580.

[Hair et al. 2021a] Hair, A., Ballard, K. J., Markoulli, C., Monroe, P., Mckechnie, J., Ahmed, B., and Gutierrez-Osuna, R. (2021a). A longitudinal evaluation of tablet-based child speech therapy with apraxia world. *ACM Transactions on Accessible Computing (TACCESS)*, 14(1):1–26.

[Hair et al. 2021b] Hair, A., Zhao, G., Ahmed, B., Ballard, K. J., and Gutierrez-Osuna, R. (2021b). Assessing posterior-based mispronunciation detection on field-collected recordings from child speech therapy sessions. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 1:181 – 185. Cited by: 3.

[Husni and Jamaludin 2009] Husni, H. and Jamaludin, Z. (2009). Dyslexic childrenâ€™s reading pattern as input for asr: Data, analysis, and pronunciation model. *Journal of Information and Communication Technology*, 8:1–13.

[Jain et al. 2022] Jain, R., Yiwere, M. Y., Bigioi, D., Corcoran, P., and Cucu, H. (2022). A text-to-speech pipeline, evaluation methodology, and initial fine-tuning results for child speech synthesis. *IEEE Access*, 10:47628–47642.

[Kathania et al. 2021] Kathania, H. K., Kadiri, S. R., Alku, P., and Kurimo, M. (2021). Spectral modification for recognition of children's speech undermismatched conditions. In Dobnik, S. and Øvrelid, L., editors, *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 94–100, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

[Kitchenham and Charters 2007] Kitchenham, B. and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. Technical report, Software Engineering Group, Keele University and Department of Computer Science, University of Durham.

[Kothalkar et al. 2021] Kothalkar, P. V., Datla, S., Dutta, S., Hansen, J. H., Seven, Y., Irvin, D., and Buzhardt, J. (2021). Measuring frequency of child-directed wh-question words for alternate preschool locations using speech recognition and location tracking technologies. In *Companion Publication of the 2021 International Conference on Multimodal Interaction*, pages 414–418.

[Li et al. 2015] Li, J., Deng, L., Haeb-Umbach, R., and Gong, Y. (2015). Robust automatic speech recognition: a bridge to practical applications.

[Li et al. 2022] Li, J. et al. (2022). Recent advances in end-to-end automatic speech recognition. *APSIPA Transactions on Signal and Information Processing*, 11(1).

[Liberman 1989] Liberman, A. M. (1989). Reading is hard just because listening is easy. *Brain and reading*, pages 197–205.

[Lileikyte et al. 2022] Lileikyte, R., Irvin, D., and Hansen, J. H. (2022). Assessing child communication engagement and statistical speech patterns for american english via speech recognition in naturalistic active learning spaces. *Speech Communication*, 140:98 – 108. Cited by: 8; All Open Access, Bronze Open Access.

[Lyakso et al. 2015] Lyakso, E., Frolova, O., Dmitrieva, E., Grigorev, A., Kaya, H., Salah, A. A., and Karpov, A. (2015). Emochildru: emotional child russian speech corpus. In *Speech and Computer: 17th International Conference, SPECOM 2015, Athens, Greece, September 20-24, 2015, Proceedings 17*, pages 144–152. Springer.

[Misra et al. 2021] Misra, A., Loukina, A., Beigman Klebanov, B., Gyawali, B., and Zechner, K. (2021). A good start is half the battle won: Unsupervised pre-training for low resource children's speech recognition for an interactive reading companion. *Lecture*

Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12748 LNAI:306 – 317. Cited by: 0.

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

[Neri et al. 2003] Neri, A., Cucchiarini, C., and Strik, H. (2003). Automatic speech recognition for second language learning: How and why it actually works. *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS)*, 5:1157–1160.

[Ngo et al. 2024] Ngo, T. T.-N., Chen, H. H.-J., and Lai, K. K.-W. (2024). The effectiveness of automatic speech recognition in esl/efl pronunciation: A meta-analysis. *ReCALL*, 36(1):4–21.

[Potamianos and Narayanan 2003] Potamianos, A. and Narayanan, S. (2003). Robust recognition of children's speech. *IEEE Transactions on speech and audio processing*, 11(6):603–616.

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

[Shahin et al. 2022] Shahin, M., Ahmed, B., and Epps, J. (2022). Speaker-and age-invariant training for child acoustic modeling using adversarial multi-task learning. *arXiv preprint arXiv:2210.10231*.

[Shinohara 2016] Shinohara, Y. (2016). Adversarial multi-task learning of deep neural networks for robust speech recognition. In *Interspeech*, pages 2369–2372. San Francisco, CA, USA.

[Southwell et al. 2022] Southwell, R., Pugh, S., Perkoff, E. M., Clevenger, C., Bush, J. B., Lieber, R., Ward, W., Foltz, P., and D'Mello, S. (2022). Challenges and feasibility of automatic speech recognition for modeling student collaborative discourse in classrooms. *International Educational Data Mining Society*.

[Taniya et al. 2020] Taniya, Bhardwaj, V., and Kadyan, V. (2020). Deep neural network trained punjabi children speech recognition system using kaldi toolkit. In *2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*, pages 374–378.

[UNICEF 2017] UNICEF (2017). The state of the world's children 2017: Children in a digital world. UNICEF.

[Watanabe et al. 2018] Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Soplin, N. E. Y., Heymann, J., Wiesner, M., Chen, N., et al. (2018). Espnet: End-to-end speech processing toolkit. *arXiv preprint arXiv:1804.00015*.

[Wilks et al. 2010] Wilks, T., Gerber, R. J., and Erdie-Lalena, C. (2010). Developmental milestones: cognitive development. *Pediatrics in Review*, 31(9):364–367.

[Yeung et al. 2021] Yeung, G., Fan, R., and Alwan, A. (2021). Fundamental frequency feature normalization and data augmentation for child speech recognition. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2021-June:6993 – 6997. Cited by: 16; All Open Access, Green Open Access.