applied sciences

MDPI

Review

Automatic Speech Recognition (ASR) Systems for Children: A Systematic Literature Review

Vivek Bhardwaj [1], Mohamed Tahar Ben Othman [2,*], Vinay Kukreja [3,*], Youcef Belkhier [4], Mohit Bajaj [5], B. Srikanth Goud [6], Ateeq Ur Rehman [7,8], Muhammad Shafiq [9,*] and Habib Hamam [8,10,11,12]

1 School of Computer Science and Engineering, Lovely Professional University, Jalandhar 144411, India; vivek.bhardwaj@outlook.in
2 Department of Computer Science, College of Computer, Qassim University, Buraydah 51452, Saudi Arabia
3 Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura 140401, India
4 Laboratoire de Technologie Industrielle et de l'Information (LTII), Faculté de Technologie, Université de Bejaia, Bejaia 06000, Algeria; belkhieryoucef@outlook.fr
5 Department of Electrical and Electronics Engineering, National Institute of Technology, Delhi 110040, India; mohitbajaj@nitdelhi.ac.in
6 Department of Electrical and Electronics Engineering, Anurag College of Engineering, Ghatkesar 501301, India; srikanth.b@anuraghyd.ac.in
7 College of Internet of Things (IoT) Engineering, Changzhou Campus, Hohai University (HHU), Changzhou 213022, China; ateqrehman@gmail.com
8 Faculty of Engineering, Uni de Moncton, Moncton, NB E1A3E9, Canada; habib.hamam@umoncton.ca
9 Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, Korea
10 Spectrum of Knowledge Production & Skills Development, Sfax 3027, Tunisia
11 International Institute of Technology and Management, Libreville BP1989, Gabon
12 Department of Electrical and Electronic Engineering Science, School of Electrical Engineering, University of Johannesburg, Johannesburg 2006, South Africa
* Correspondence: maathaman@qu.edu.sa (M.T.B.O.); vinay.kukreja@chitkara.edu.in (V.K.); shafiq@ynu.ac.kr (M.S.)

Abstract: Automatic speech recognition (ASR) is one of the ways used to transform acoustic speech signals into text. Over the last few decades, an enormous amount of research work has been done in the research area of speech recognition (SR). However, most studies have focused on building ASR systems based on adult speech. The recognition of children's speech was neglected for some time, which means that the field of children's SR research is wide open. Children's SR is a challenging task due to the large variations in children's articulatory, acoustic, physical, and linguistic characteristics compared to adult speech. Thus, the field became a very attractive area of research and it is important to understand where the main center of attention is, and what are the most widely used methods for extracting acoustic features, various acoustic models, speech datasets, the SR toolkits used during the recognition process, and so on. ASR systems or interfaces are extensively used and integrated into various real-life applications, such as search engines, the healthcare industry, biometric analysis, car systems, the military, aids for people with disabilities, and mobile devices. A systematic literature review (SLR) is presented in this work by extracting the relevant information from 76 research papers published from 2009 to 2020 in the field of ASR for children. The objective of this review is to throw light on the trends of research in children's speech recognition and analyze the potential of trending techniques to recognize children's speech.

Keywords: automatic speech recognition; MFCC; children's speech recognition; acoustic model; systematic literature review (SLR)

check for updates

Citation: Bhardwaj, V.; Ben Othman, M.T.; Kukreja, V.; Belkhier, Y.; Bajaj, M.; Goud, B.S.; Rehman, A.U.; Shafiq, M.; Hamam, H. Automatic Speech Recognition (ASR) Systems for Children: A Systematic Literature Review. *Appl. Sci.* **2022**, *12*, 4419. https://doi.org/10.3390/app12094419

Academic Editor: Lijiang Chen

Received: 21 March 2022
Accepted: 23 April 2022
Published: 27 April 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

In recent decades, remarkable progress has been accomplished in developing functional spoken dialog systems and automatic speech recognition (ASR) systems, and both are utilized in different applications. Several techniques have been proposed by researchers

to improve the performance and recognition accuracy of ASR systems but these are mainly focused on speech recognition in adult speakers. According to the studies, the ASR system's efficiency is lower when it is tested using children's speech; this finding directed attention toward the area of more robust ASR systems for interpreting children's utterances. Introducing suitable amounts of children's speech data for training the system is one way of improving children's speech recognition. However, the majority of public data sets are compiled with the assistance of adult speakers. Collecting a children's speech corpus for training the ASR system is difficult and the data sets are usually smaller than the adult corpus. Acoustics and linguistic properties such as the spectral and temporal features of adults and children are also different. As a result of the variance in these characteristics, there is a mismatch between children's and adult speech. The main reason behind these differences is the morphological and anatomical variabilities in the vocal tract and the fact that children have less control over prosodic features such as pitch, power, rhythm, and intonation. Various speaker normalization and adaptation techniques have been proposed to date, to reduce the mismatches. Thus, another method to improve children's speech recognition is by reducing the size of the acoustic mismatch between adults' and children's speech by applying different algorithms. Thus, due to a limited data set and differences in acoustic and linguistic properties, recognizing children's speech remains one of the most challenging parts of the ASR system. In the current research study, we have conducted an SLR exploring the studies published in the field of children's speech recognition. An SLR is conducted by following the steps used by Asad Ali and Carmine Gravino in [1]. Using a block diagram, Figure 1 depicts the steps of developing a children's speech recognition system in various matched and mismatched acoustic situations.

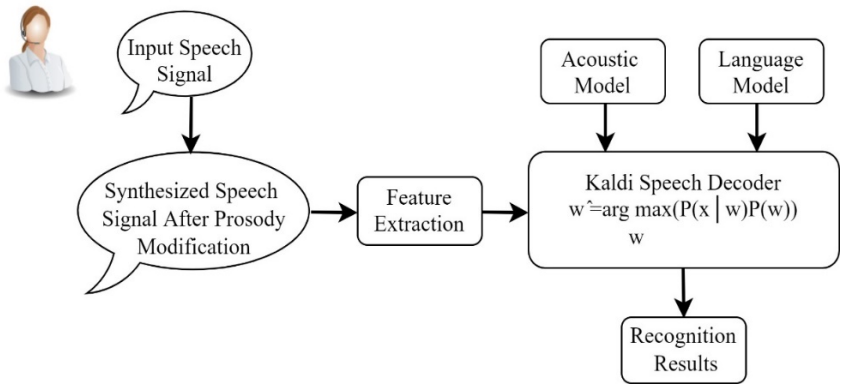

Figure 1. ASR model for recognizing children's speech with acoustic variabilities.

In contrast to earlier research, this systematic literature review offers a detailed examination of children's speech recognition in the wider field of ASR. Furthermore, past research has focused on adult speech recognition and no SLR has been undertaken for the recognition of children's speech in a variety of acoustic environments. Therefore, in this SLR, we began by providing a review of ASR, its categories, speech corpora, and toolkits for recognizing children's speech, which is the study's main topic. Following that, we discussed speech recognition issues and identified several methods for improving voice recognition in children and a variety of application areas. This provides a solid theoretical framework for the reader to build upon in order to fully appreciate the topic at hand, which is children's speech recognition.

In total, 76 publications were included in the SLR, which were published throughout an 11-year period between 2009 and 2020. The following are some of the details gleaned from the above papers:

(1) The several ways of extracting front-end features.
(2) The various strategies for acoustic modeling.
(3) The audio corpora on which the system was trained and tested.

(4) The various types of speech recognition toolkits on the market.
(5) The various natural languages with which the algorithm was trained and tested.
(6) The types of environmental conditions (noisy, quiet, or neutral).
(7) The techniques for evaluating the system's performance.
(8) The publication's type (journal, conference, or workshop).
(9) The conference, workshop, or journal where the paper was presented and published.
(10) The distribution of papers over time.

The following points are the numerous gaps that were identified, based on the essential works chosen for this study and the information gathered about them.

(1) Continuous and spontaneous speech recognition for child speakers has received less attention than isolated and connected word speech recognition.
(2) Sufficiently large speech databases in numerous regional languages are not publicly available.
(3) Speech recognition systems' efficiency and accuracy are quite low in regional languages with mismatched acoustic circumstances.
(4) The implementation of hybrid feature extraction approaches for obtaining acoustic features.

Motivation: Speech is an effective way for a human and a machine to communicate. Speech recognition, which is part of the computational linguistics field, is used to convert a received speech signal into text. ASR systems have only been developed for a few of the 7000 languages spoken on the globe. The recent changes in schools attest to the fact that the last 18 months have transformed the globe in ways we could never have foreseen. While it was once believed that e-learning would play a significant role in education in the future, it was utterly unforeseen that it would completely replace traditional classroom instruction so quickly. As a result of the COVID-19 pandemic, video conferencing has become an essential teaching tool, whether through virtual tutoring, language applications, video conferencing technologies, or online learning software. By delivering real-time captioning for all students, including those who are deaf or hard of hearing, speech recognition technology can increase the accessibility of video platforms, making online learning more inclusive. Additionally, students who do not speak English as a first language may find it easier to read captions and transcripts than to follow the teacher's voice, which may have an unfamiliar accent or dialect. Speech recognition technology is also beneficial for children who have difficulty with handwriting or spelling. Instead of typing or writing by hand, children with dyslexia and other learning impairments can utilize speech-to-text technology to write using their voices. Children make up a sizable segment of the market that will benefit from advances in multi-media equipment that uses ASR technology. Children are one of the most likely users of computers for social interaction in terms of multimedia games, educational software, and education materials. In line with this finding, children are often more comfortable and joyful while using spoken-language interfaces. ASR interfaces must first recognize and then adapt to the user's language, to match with or complement the user's speech in order to make the machine more interesting to interact with. Thus, an ASR system for children is essential to make human–machine interaction more flexible. Abbreviations part contains a list of abbreviations used in this research.

This systematic review is organized as follows: Section 2 highlights current speech technology, ASR, the speech corpora available, ASR toolkits, challenges, children's vs. adults' speech patterns, and various approaches for the better recognition of children's speech. The methodology we have followed to conduct the SLR is described in Section 3. The results obtained for each research question are presented in Section 4. Finally, Section 5 concludes with all the outcomes of conducting this SLR.

2. Background

2.1. Speech Technology

Speech technology involves the processing of spoken words; the words are treated either as a speech signal or as a natural language. In general, this refers to the study

and processing methods of speech signals. To process the speech signals, signals are first converted into a digital format, so this is a special case of digital signal processing. Speech technology is a multi-disciplinary field that includes linguistics, psychology, signal processing, acoustics, pattern recognition, AI, and ML. Major research areas that come under the umbrella of speech technology are:

- Speech recognition (speech to text);
- Speech synthesis (text to speech);
- Speaker recognition (identification of the speaker);
- Speech encoding (compression of speech);
- Multimodal interaction (modeling of natural text).

The purpose of this paper is to review and analyze the available studies on children's ASR. To deal with resource availability, such as the speech corpus, a variety of ASR technologies have evolved into diverse language-specific systems.

2.2. Automatic Speech Recognition

Speech recognition is a multidisciplinary research area consisting of linguistics and computer science. It is also called speech-to-text (STT), computer speech recognition, or ASR. The ASR system accepts spoken words in an audio format, such as .raw or .wav files, then generates its content in text format and makes a computer understand the natural language. By "understand", we mean to react appropriately, or transform the speech signal into another form, i.e., text. ASR systems generate models of given training data and speech is tested using these models, trained during the training phase. Depending on the type of speech, speaker, vocabulary, and environmental conditions, ASR systems are classified into different categories [2], which are shown in Figure 2.

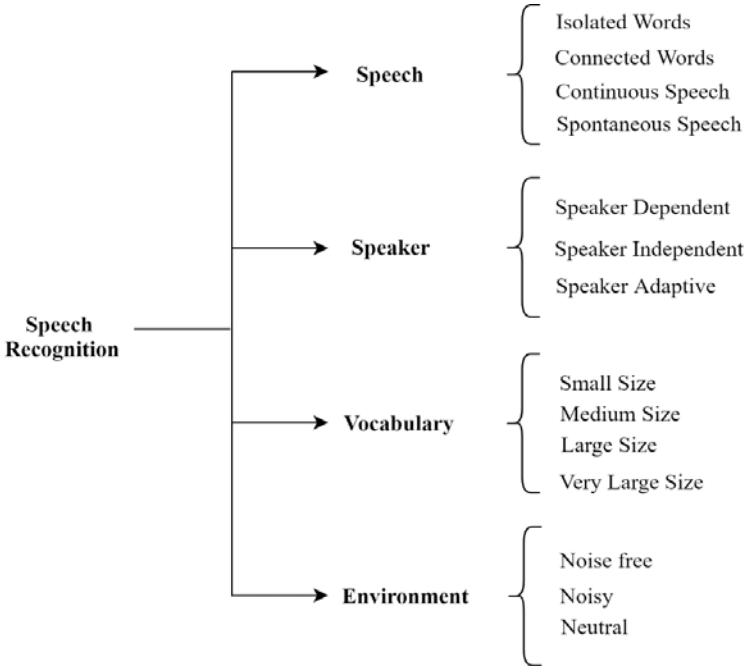

Figure 2. ASR system categories.

An isolated word speech recognition system is one in which the software operates on one word at a time and requires a pause between each utterance, e.g., "one", "five", "Punjabi". A connected word speech recognition system is similar to an isolated word system, but it allows different words or utterances to be spoken together, with a small pause between them, e.g., "7,814,108,877". A continuous ASR system runs on speech signals in which different words are connected together without any pauses, e.g., "Today is Sunday". Spontaneous word speech recognition, in which the system operates on natural speech

and the system accepts every type of input, which may have grammatical errors, might be within a noisy environment, or may be false statements, e.g., made by commentators.

A speaker-dependent system must be trained for a specific speaker and learns speaker voice characteristics through training. Conversely, a speaker-independent system does not require any training and can be used by different people without any previous training to recognize each speaker's speech characteristics. The speaker-adaptive system uses a small amount of speaker-specific data to adapt a speaker-independent ASR system, such that the system is attuned to the speech of a specific speaker. The vocabulary or dictionary consists of words that can be recognized by the ASR system. Thus, the size of its vocabulary also affects the accuracy of the speech recognition system. Tens of words make up a narrow- or small-vocabulary speech recognition system; a medium-sized vocabulary speech recognition system consists of hundreds of words. A large-sized vocabulary speech recognition system consists of thousands of words, whereas a very large-sized vocabulary speech recognition system consists of tens of thousands of words. Larger-sized vocabularies make it more difficult to recognize specific words compared to small-sized vocabularies.

ASR systems can be tested under different environmental conditions, such as noisy, noise-free, or neutral environments. Therefore, environmental conditions play an important role in developing a robust ASR system.

2.3. Resources for Automatic Speech Recognition

2.3.1. Speech Corpora

A speech corpora or corpus is a collection of spoken audio files, along with text transcriptions of the audio files [2]. In the field of speech recognition, speech corpora are used for acoustic analysis and to build acoustic models for recognizing the speech pattern or speaker. The availability of a sufficient amount of public speech corpora or corpus plays an important role when researching the field of speech recognition. While doing research in SR, some researchers use their own private speech corpus, whereas others prefer publicly available speech corpora. Most of the speech corpora used for children's speech recognition were developed with the assistance of the 5–18 years age group. Some of the most commonly used, publicly available children's speech corpora for acoustic modeling [3] in the different languages of British English, Russian, and Italian are:

- PF-STAR children's speech corpora;
- TIDIGITS;
- ChildIt corpora;
- CID children's speech corpora;
- CUKids;
- CMU Kids corpora;
- EmoChildRu.

Adult speech corpora were also required for training the ASR system to achieve a better recognition rate when the system was tested under mismatched acoustic conditions. A few of the popular speech corpuses used for training children's speech recognition systems [3] are:

- WSJCAM0;
- TED-LIUM ASR corpus;
- TIMIT.

2.3.2. ASR Toolkits

In the case of general, operational, and customer-facing speech recognition, users may prefer to use products that are available commercially, such as Cortana, Google Cloud Speech API, Amazon Lex, or Dragon. However, in the case of research and development (R&D), when developing ASR applications, researchers have several choices of open-source speech recognition toolkits for building a speech recognition system. A list of top

contenders in terms of free or open-source toolkits for speech recognition is shown in Table 1.

Table 1. Open-source ASR toolkits.

Speech Recognition Toolkit	Description	License	Open-Source	Programming Language	Supported Language
Hidden Markov Toolkit (HTK)	HMM neural net	HTK specific	Not strictly open source	C	English; version 3.5 released December 2015
CMU Sphinx	HMM	Berkeley Software Distribution (BSD)	Yes	Java	English, German, Russian, Mandarin, French
Kaldi	Neural net, finite-state transducers	Apache v2.0	Yes	C++	English
Julius	HMM trigrams	Berkeley Software Distribution (BSD)	Yes	C	Japanese, English

The hidden Markov toolkit (HTK) is a freely available and portable speech recognition toolkit for use by researchers. The Speech, Vision and Robotics group at Cambridge University used the C programming language to create the toolkit. The HTK is used for building large-vocabulary ASR systems using HMMs; along with this application, it is used in many other contexts, such as character recognition, speech synthesis, and DNA sequencing. From the year 2000, the HTK became freely available in the form of source code and is accessible through htk.eng.cam.ac.uk as a speech recognition research platform [4].

CMUSphinx is one of the more popular speech recognition toolkits, which contains multiple tools for creating speech applications. It was written and developed by researchers in Java programming at Carnegie Mellon University. Some components of the toolkit are Pocketsphinx, Sphinxbase, Sphinx4, and Sphinxtrain [5]. In 2000, the Sphinx group made some components of their speech recognizer open-source.

Kaldi [6] is another open-source SR toolkit written in the C++ programming language and is freely available using the Apache License v2.0, making it easily accessible for a wide community of researchers. Kaldi came into existence in 2009 at Johns Hopkins University via a workshop. It uses finite-state transducers for building speech recognition systems and supports deep neural networks, MMI, boosted MMI, feature-space discriminative training, and MCE discriminative training.

Julius is a high-performance speech recognizer toolkit with a large vocabulary and a continuous speech recognition (LVCSR) decoder and is freely available for researchers and developers working in speech technology. It is based on context-dependent HMM and word N-gram models. Julius is a free software toolkit developed by the IPA Japanese dictation toolkit project, which has been researching Japanese LVCSR since 1997 [7].

2.4. Challenges in the Development of the ASR System

There is great variability in the different speech dimensions. Therefore, this makes speech recognition a challenging task. However, research in this field is moving forward in terms of speech recognition. The various problems researchers are facing while processing the speech signals into the text are as follows.

2.4.1. Environment and Channel Characteristics

Background noise, room acoustics, and microphone properties play an important role in speech recognition. Voice-recording devices detect acoustic waveforms generated from speech by articulated sounds. Background noises in open spaces and areas like the underground, railway stations, and large rooms make it very difficult for ASR systems to understand and differentiate the specific soundwaves of speech from the anchor-voice microphone properties.

2.4.2. Echo

In acoustic and speech signal processing, echoes are sound waveforms that arrive at the listener after being reflected across various surfaces, such as tables, walls from a building, or other furniture. Due to this effect, sound waveforms received by the receptors are captured with less clarity and this reduces the accuracy of recognition.

2.4.3. Speaking Style

During the speech recognition process, the speaker's speaking style is very significant. What type of speech the speaker is using, such as spontaneous speech or continuous natural speech, is important since these types are difficult to recognize, compared to the isolated or connected speech recognition forms.

2.4.4. Speaker Characteristics

The rate of speech, interpreting prosodic features such as pitch, power, intonation, stress, speaker age, and variations in pronunciation even when the same word is spoken by the same speaker also affect the recognition accuracy. Another major factor that makes it difficult to recognize and interpret speech is the variety of accents in languages. If the same words can be pronounced in multiple ways, the phonetics and syllables of the word differ for each spoken word, which makes it difficult to recognize for the ASR system.

2.5. Children's vs. Adults' Speech

The variations in the acoustics and linguistics of children's and adults' speech are discussed in this section. With increasing age and physical developmental changes in children, the spectral and temporal properties of speech are continuously affected [8,9]. As a result, the characteristics of the adults' speech diverge from those of children as time passes. The main reason behind these differences is the morphological and anatomical differences in the vocal tract and the child's lesser control of prosodic features such as pitch, power, rhythm, and intonation. Children's formant and fundamental frequencies are higher in comparison to adults because of the smaller vocal folds and shorter vocal tract [10,11]. Age-dependent changes were found in the formant and fundamental frequencies of a child speaker aged between three and thirteen, in these studies. When children are young, they have very little experience in articulating sounds and have not learned as many words of the language compared to adults. Thus, the vocabulary size of the children is small and is different from adults since children use their associative skills and imagination to invent their own words. In general, with increasing age, the ability of children to efficiently use a language improves. There is a decrease in disfluencies with age; children's voices reach adult-speaking levels at about 12 to 13 years of age. In the case of children between 8 and 10 years, the probability of the mispronunciation of words is high, almost twice that in 11- to 14-year-old children.

2.6. Approaches for the Better Recognition of Children's Speech

Many studies confirmed that research has been conducted into targeting the adaptations of adult speech recognition systems toward children's speech recognition. In the above section, we have discussed the various acoustic differences that exist between children's and adults' speech, which will be an important area to investigate when moving forward. A number of researchers have worked on compensating for the various acoustic variations influenced by short vocal tract length in children using vocal tract length normalization (VTLN). Due to these variations, an ASR system trained using adult speech delivered a poor recognition rate when tested using children's speech and mismatched acoustic conditions. Thus, in this section, we discuss the approaches used for better children's SR under mismatched conditions.

2.6.1. Speaker Normalization: Vocal Tract Length Normalization (VTLN)

Speaker normalization is used to normalize the acoustic data to minimize the mismatch with acoustic models. VTLN is one of the techniques for rapid speaker normalization or adaptation that is broadly used in speech recognition. It aims at narrowing down the inter-speaker acoustic irregularities due to differences in the vocal tract length (VTL) by warping the frequency axis of the speech spectrum of speakers. Research performed using VTLN shows a better recognition rate when the ASR system was trained using adult speech and tested using children's speech [10,12,13]. Results obtained with the VTLN are still not satisfactory because there are other factors, along with differences in the VTL, that make children's speech different from adult speech. Along with VTLM, the model-based acoustic adaptation technique of HMM/GMM parameters and Gaussian parameters, such as maximum a posteriori (MAP) and maximum likelihood linear regression adaptation (MLLR) are also helping to increase the recognition rate of the ASR system.

2.6.2. Maximum a Posteriori (MAP)

MAP [14–16], also called Bayesian adaptation, is one of the most popular model-based acoustic model adaptation systems. MAP estimates the model parameters more robustly compared to maximum-likelihood (ML) estimation when data availability is lower because its estimates do not require a massive dataset of speech samples as it also makes use of a priori information with ML estimation. The re-estimation of model parameters is performed independently of one another in the MAP adaptation. The MAP estimate is then set up by moving the values of the original prior parameter toward the ML estimates. Thus, you can say that the MAP estimate is a weighted average between the prior estimate and ML estimate.

2.6.3. Maximum Likelihood Linear Regression (MLLR) Adaptation

MLLR is a model-based technique for GMM-based systems [17,18]. The model parameters are adapted indirectly, without any updating, by learning linear transforms to adapt the acoustic classes, i.e., mean and covariance parameters. The main advantage of the adaptation methods based on a linear transform is that all Gaussian parameters can be adapted with only a few or a single transformation. For each acoustic class, there is a separate transformation. The acoustic class can be chosen as a subset of phonemes, an individual speaker, or an acoustic environment. MLLR has been demonstrated to work consistently well, down to just 10 s of adaptation data.

Constrained MLLR (cMLLR): the cMLLR is a version of MLLR in which the same linear transform is used for the mean and the covariance parameters. This is very interesting because it also corresponds to a linear transform of the features. This means that given a GMM system, a cMLLR adaptation transform can be estimated and used as a feature space adaptation transform (speaker normalization) for any acoustic modeling approach (e.g., a neural network). In signal processing, feature space MLLR (fMLLR) is a global feature transform that is particularly applied in a speaker-adaptive way, wherein fMLLR mutates the acoustic features of the signal to speaker-adapted acoustic features with the help of a transformation matrix multiplication operation. In some studies, fMLLR has also been called cMLLR [19].

2.6.4. Speaker-Adaptive Training (SAT)

In acoustic modeling, SAT is a standard and well-established technique for GMM models. Acoustic models trained with the help of the SAT adaptation technique do not depend on the speakers being present during the training and generalize better to obscure the speakers who are present during the testing. In SAT adaptation, transforms are computed at training time as well as at test time. This has the advantage of consistency, as it assumes that an adaptation transformation is learned for every speaker encountered by the system [20,21].

2.6.5. Dynamic Time Warping (DTW)

To recognize the compatibility of a speech signal, a special algorithm is needed, that of dynamic time warping (DTW). When comparing the similarity of a pattern across time zones, the DTW approach is utilized [22]. The narrower the distance formed, the closer the two sound patterns are. Because their sound patterns are identical, the two voices are assumed to be the same. The initial data from the voice recognition process are then turned into frequency waves.

2.7. Significance of ASR in Various Fields

Speech applications are installed publicly in railways, multiplexes, airport communication, and tourist places where people are answered when they pose instant queries. This technology plays an important role in leveling the gap between people with language communication problems. Because it is difficult for blind and sight-impaired individuals to read from a screen and writing on a sheet of paper every time is inefficient in today's world, speech recognition is a potential interface for physically handicapped persons. Major real-life applications where speech recognition plays an important role are:

- Car or vehicle infotainment systems;
- In education and daily life, such as talking to robots;
- Telephony and computer gaming;
- Controlling digital devices (Alexa and Google Home);
- Health care (medical documentation and therapeutic use);
- In the military (aircraft and helicopters);
- People with disabilities;
- Intelligent virtual assistants (IVAs).

3. Methodology

This survey is based on the guidelines and steps presented by Ali and Gravino in the SLR [1].

3.1. Research Question (RQ)

Before starting a study, research professionals, as well as successful student researchers, develop research questions because they are key to conducting the SLR research process. These questions influence the remaining steps in the research process by outlining exactly what the authors are aiming to learn from the studies. The main goal of this SLR is to identify and explore the research papers and research work conducted in the field of children's speech recognition. The following RQ were identified to direct this review:

RQ 1: What are the different categories of research papers that were included in the review process?
RQ 2: How were front-end acoustic features extracted from the speech signal?
RQ 3: Which acoustic models have been used for training and testing in the studies?
RQ 4: What type of speech datasets were used during the training and testing of the models in the papers included in the SLR?
RQ 5: Which toolkits are used by the authors for speech recognition?
RQ 6: What are the different natural languages used to develop the ASR system?
RQ 7: What type of environmental conditions were used by researchers to develop the speech recognition system?
RQ 8: Which are the evaluation methods used to assess the overall performance?

3.2. Search Scheme

After identifying the research questions, we identified and downloaded the research papers, based on the following search criteria:

- Primary search;
- Secondary search.

For conducting the primary search, the following steps were followed:

1. Identification of the key terms related to the research questions.
2. Synonyms of the key terms used in step 1 are also used.
3. Limitation of the search results, where the Boolean operators AND and OR are used during the searching.

The research papers that were missed during the primary search were identified by reviewing the references of the primary search papers in a secondary search.

The key terms used in the primary search string were:

- ("Children speech recognition");
- ("Automatic speech recognition") AND ("children");
- ("speech recognition" OR "children speech recognition") AND ("Children" OR "children's").

The following databases were used for searching and selecting the research papers for conducting this review:

- IEEE Xplore;
- Google Scholar;
- ScienceDirect;
- ResearchGate;
- Scopus;
- Springer.

The research papers used in the review were published from 2009 to 2020. The above-listed databases were used to extract the research papers. During the primary search, we identified and selected about 165 research papers. After that, a secondary search was conducted to identify the missed research papers in the initial search, with the help of the references in the selected papers. By using a secondary search, we found 20 more relevant research papers. Thus, after combining both search results, a total of 185 papers were selected. Out of these 185 papers, eventually, 76 publications were used for conducting the systemic review by applying various inclusion/exclusion and quality assessment criteria. The inclusion/exclusion criteria used to filter the papers are explained in the next section.

3.3. Study Selection

Once the searching and selection of the research papers were conducted based on the titles and abstracts, inclusion/exclusion and quality assessment criteria were applied to gain more authentic and relevant research papers. The inclusion and exclusion criteria used for filtering and deciding which research papers were to be included are the following:

- Inclusion Criteria:
    1. Research papers that used only a children's or children + adults speech corpus for building the ASR system.
    2. Research papers that focus on children's speech recognition in terms of speech technology.
    3. The research papers that are published in conferences, as well as journals; only the papers that were published in the journal were included in the survey.
- Exclusion Criteria:
    1. Research papers that were related to speech and that focused on speech recognition but did not use children's speech for building the speech recognition system.
    2. Research papers that used children's speech in an area of speech technology other than speech recognition.
    3. All the duplicated research papers that were downloaded from various digital databases.
    4. Review papers.

3.4. Quality Assessment Benchmarks

After defining the inclusion and exclusion criteria for the research papers, quality assessment benchmarks are the final step we used to ascertain the reliability and applicability of the research papers in the SLR. We can, therefore, say that quality assessment benchmarks can be considered the additional criterion for the selection of the research papers. Different quality assessment rules were applied to evaluate the quality of the research papers. Ten quality assessment questions were used to establish the final list of research papers for the SLR. Depending upon their quality and ability to answer the RQ, we have decided to give scores for each question, based on the following rules.

Rule 1: A score of 1 is given if the answer to the RQ meets the full requirements.
Rule 2: A score of 0 is given if RQ is not answered.
Rule 3: A score of 0.5 is given if the answer to the RQ is average.
Rule 4: A score of 0.75 is given if the RQ answer is above average.
Rule 5: A score of 0.25 is given if the RQ answer is below average.

The below-mentioned questions are used to determine the quality assessment criteria.

Q 1. Are the objectives of the research work are clearly defined?
Q 2. Is the area of speech recognition used clearly defined or not?
Q 2. Are the research gap and challenges mentioned in the papers?
Q 3. Are the feature extraction methods deliberated and well-defined?
Q 4. Are the classification methods and acoustic units clearly stated?
Q 5. Are the details of the speech corpora clearly defined and the experiments performed on sufficient speech corpora?
Q 6. Are the language and environmental conditions defined in the papers?
Q 7. Are the evaluation methods used to evaluate the performance appropriately?
Q 8. Is the experimental work conducted or not?
Q 9. How recent are the research papers?
Q 10. Are the limitations of the conducted research work analyzed explicitly?

After applying the quality assessment rules, the total score was counted by adding the score of each question. If the total score is more than 6 then the paper was included in the SLR and the rest of the papers were excluded. Finally, we have selected the 76 most punctilious and relevant studies after conducting the steps of inclusion, exclusion criteria, quality assessment questions, and removing the duplicate research papers.

3.5. Data Extraction

In this section, the final 76 research papers are used to extract the data to answer the research questions mentioned in Section 3.1. For each paper, we have assigned a paper ID for easy identification, listed at the end of the paper in Appendix A. For all the papers published in the year 2020, the authors assigned a paper ID as A1, A2, . . . , AN, and for the year 2019 B1, B2, . . . , BN. In this way, we have assigned unique IDs to all the 76 research papers included in the SLR. During the data extraction process, it was found that not all the research papers gave answers to all the questions. Details regarding the counts of research papers published each year from 2020 to 2009 are shown in Figure 3.

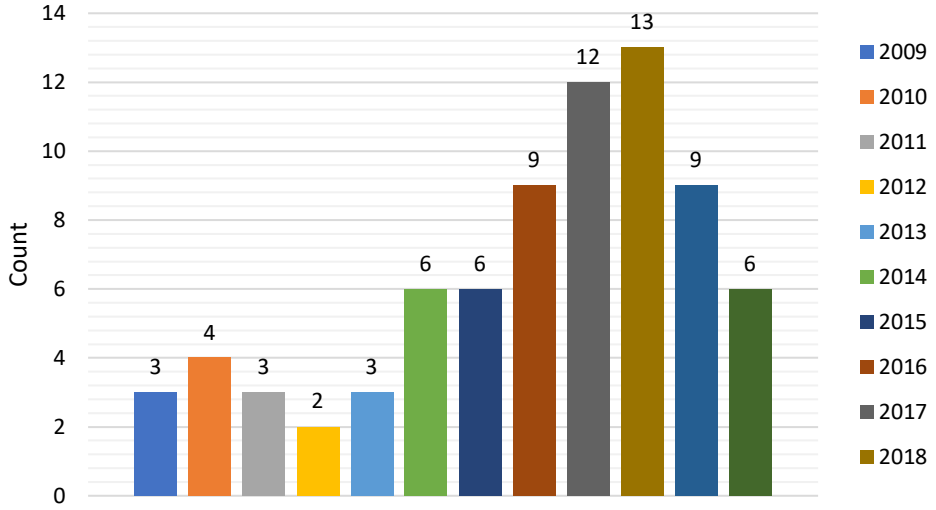

Figure 3. Year-wise publication count from 2009 to 2020.

4. Results

In this section, we discuss the results obtained for each research question mentioned in Section 4.1.

4.1. RQ 1: What Are the Different Categories of Research Papers That Were Included in the Review Process?

76 research papers are included to conduct a systematic review and fall into three categories: journal papers, conference papers, and workshop papers. Figure 4 presents the percentage distribution of research papers between these three categories. Most of the research papers included in the review were published in conference proceedings and cover 62% (47 papers) of the papers used. Then, 24% (18 papers) of the papers were published in journals and, thus, fall into the journal category. The rest of the 14% (11 papers) belong to the workshop papers group. Information regarding the publication of the papers is presented in Tables 2–4.

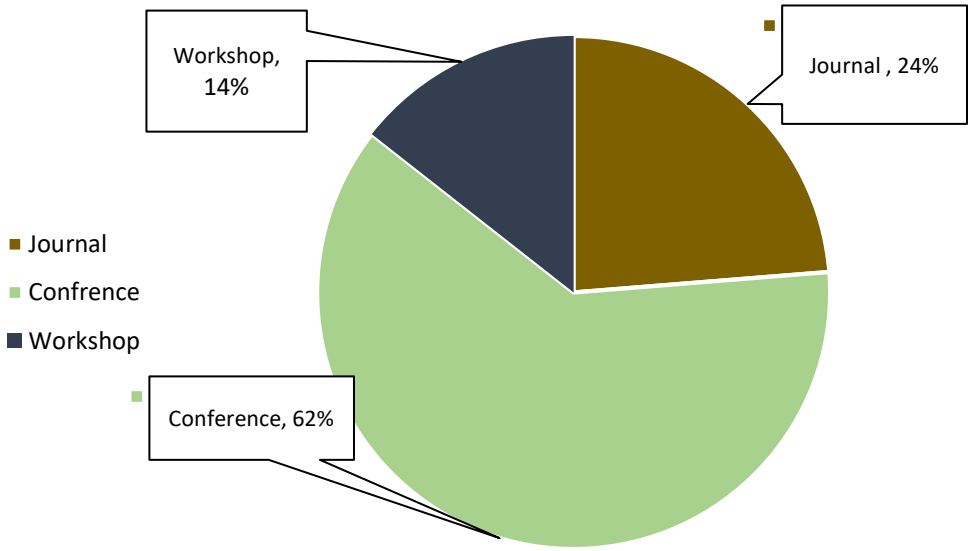

Figure 4. Percentage of journal, conference, and workshop papers included in the SLR.

Table 2. Conference research papers included in the review.

Paper ID	Conference Title	Count	%
A3, C1, C3, B7, D4	ICASSP	5	10.64
A2, C2, G4, C8, E3, I2, K1	Signal Processing and Communications (SPCOM)	7	14.89
A6	International Conference on Inventive Research in Computing Applications	1	2.13
B4, D1, D11 E7, E8, F1, F2, F3, C9, D5, D10, H3, J1, K3, L1, L2, L3	INTERSPEECH	17	36.17
C5	International Symposium on Communications and Information Technologies (ISCIT)	1	2.13
C7	Spoken Language Technologies for Under-Resourced Languages	1	2.13
D2, E6, F6	TENCON	3	6.38
C11, E4	Italian Computational Linguistics Conference (CLiC-it)	2	4.26
G1	International Student Project Conference (ICT-ISPC2014)	1	2.13
B8	International Symposium on Multimedia and Communication Technology (ISMAC)	1	2.13
B9	Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)	1	2.13
E9	International Symposium on Chinese Spoken Language Processing (ISCSLP)	1	2.13
D12	International Conference on Asian Language Processing (IALP)	1	2.13
H1	Speech and Language Technology in Education	1	2.13
I1	International Conference on Computational Intelligence, Modeling and Simulation	1	2.13
J2	International Conference on Artificial Intelligence	1	2.13
J3	Italian Association of Speech Sciences	1	2.13
H2	Indicon	1	2.13

Table 3. Workshop research papers included in the review.

Paper ID	Workshop Title	Count	%
A4	CCWC	1	9.09
B1, B3	ASRU	2	18.18
E5, G6, K2	Spoken Language Technology Workshop (SLT)	3	27.27
E2, G2, G3, G5, D7	Workshop on Child Computer Interaction (WOCCI)	5	45.45

Table 2 shows the name, count, and percentage of the papers published in the conference proceedings. From the table, it is clear that most of the conference papers, i.e., 36.17%, were published in INTERSPEECH. In addition, 14.89% of the conference papers were published in Signal Processing and Communications (SPCOM), and ICASSP published 10.64% of conference papers. Then, 6.38%, 4.26%, and 2.13% of the papers were published in TENCON, Italian Computational Linguistics Conference, and International Symposium on Chinese Spoken Language Processing (ISCSLP) conference proceedings, respectively. The rest of the papers (25%) were published in the remaining 13 conferences mentioned in Table 2.

Table 4. Journal research papers included in the review.

Paper ID	Journal Name	#	%
A1, A5, C13, D6	Computer Speech & Language	4	22.2
B2	Pattern Recognition Letters	1	5.6
B5, D9	IEEE SIGNAL PROCESSING LETTERS	2	11.1
B6, C12	Digital Signal Processing	2	11.1
C4	CSI Transactions on ICT	1	5.6
C6, C10	Circuits Syst Signal Process	2	11.1
D3	Trends in Hearing	1	5.6
D8	Frontiers in psychology	1	5.6
E1	International Journal of Computer, Electrical, Automation, Control, and Information Engineering	1	5.6
F4	The Journal of the Acoustical Society of America	1	5.6
F5	International Journal of Speech Technology	1	5.6
K4	EURASIP Journal on Audio, Speech, and Music Processing	1	5.6

Table 3 shows details of the workshop papers. The table shows that 45.45% of the papers were published in the Workshop on Child Computer Interaction (WOCCI) and 27.27% in the Spoken Language Technology (SLT) workshop. The remaining 27.28% were published in the Automatic Speech Recognition and Understanding Workshop (ASRU) and Computing and Communication Workshop and Conference (CCWC).

Finally, in Table 4, we present the details of the papers published in the journals. Computer Speech & Language journal published 22.22% of the journal category papers. IEEE Signal Processing Letters (11.11%), Digital Signal Processing (11.11%), and the CSI Transactions on ICT (11.11%) journals published 33.33% of the journal papers. The rest of the 8 journals mentioned in the table published the remaining papers.

4.2. RQ 2: How Were the Front-End Acoustic Features Extracted from the Speech Signal?

Extracting the acoustic features from the speech signal plays an important role in getting higher accuracy and performance from the ASR system. There are several feature extraction techniques available to extract features in the front end. Figure 5 shows the various feature extraction techniques used by the researchers. From the studies, it was found that 80.26% of research papers used MFCC in the front end, and it was the first choice of the researchers to extract the acoustic features. The PLPCC feature extraction technique was used in 5.26% of the papers. Spectral moment features were extracted by using SMAC in 3.96% of the papers and 2.63% of the papers used the LPC feature extraction process. The rest of the 14.47% of papers used VMD-MFCC, PNS-MFCC, BFCC, GFCC, running spectrum filtering (RSF), running spectrum analysis (RSA), PLP, stochastic feature mapping (SFM), and the PMVDR feature extraction process.

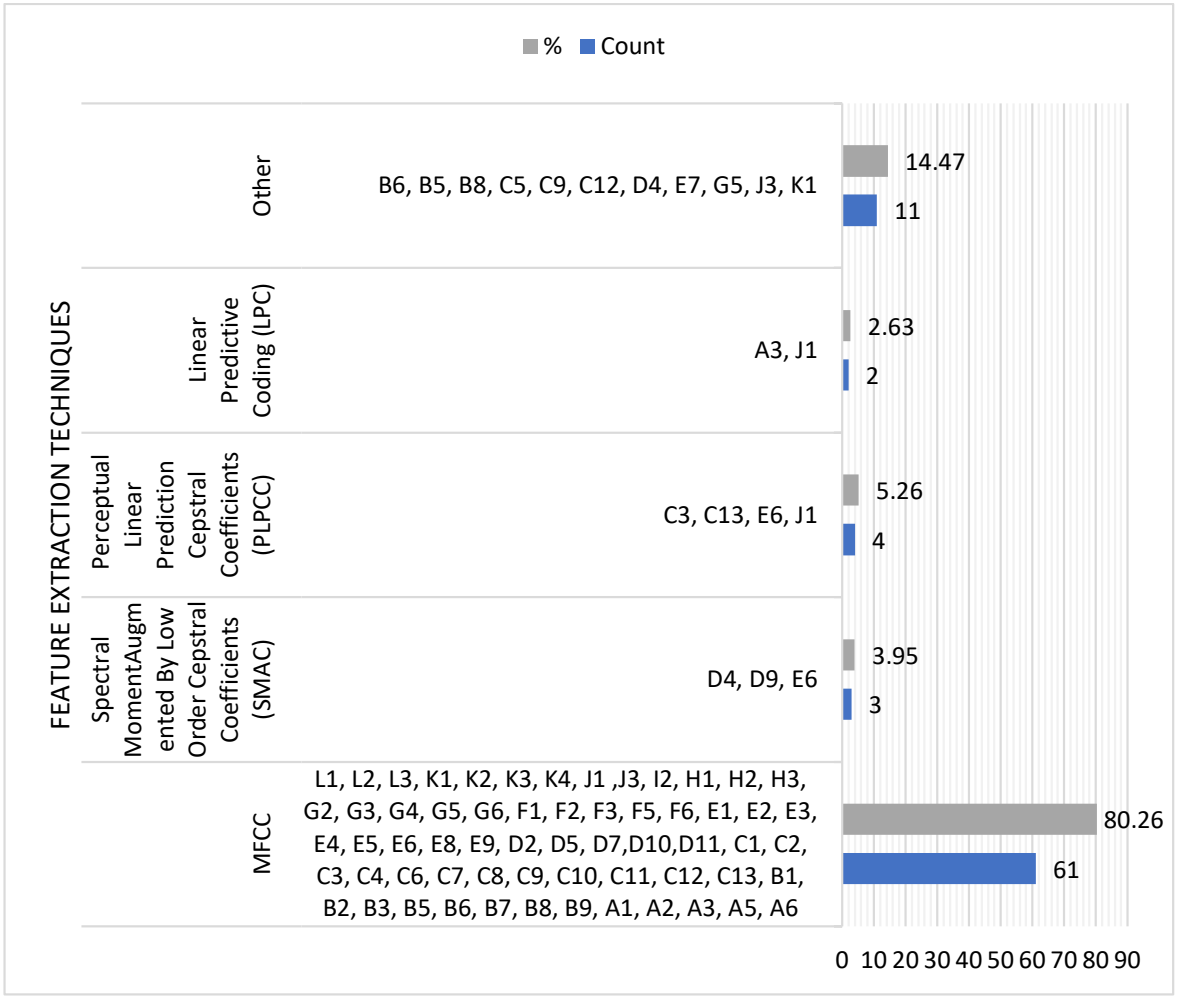

Figure 5. Techniques used to extract the acoustic features.

4.3. RQ 3: Which Acoustic Models Have Been Used for Training and Testing in the Studies?

An acoustic model is a statistical representation in ASR to show the relationship between the audio signals and phonemes that make up a word. The model is trained with the help of audio files or recordings and the corresponding text transcriptions. Figure 5 shows the different acoustic models used in the research papers. From the studies, it was found that most of the researchers trained the system using multiple hybrid acoustic models. The most commonly used models are based on GMM, SGMM, and DNN. The GMM-HMM-based acoustic model was used in 44.76% of the papers. SGMM-HMM was also used in 9.21% of papers. In addition, 35.53% used a hybrid DNN-HMM acoustic model, and 14.47% solely used DNN as an acoustic model to train the system. Some researchers used the RNN-based LSTM model, which comprise 10.53% of the papers. Papers that used HMM-based acoustic models comprise 14.47% of the papers. The rest of the papers used a time-delay neural network (TDNN), CNN-based acoustic model, Baum–Welch algorithm, Gaussian mixture model (GMM), and support vector machines (SVM), as shown in Figure 6.

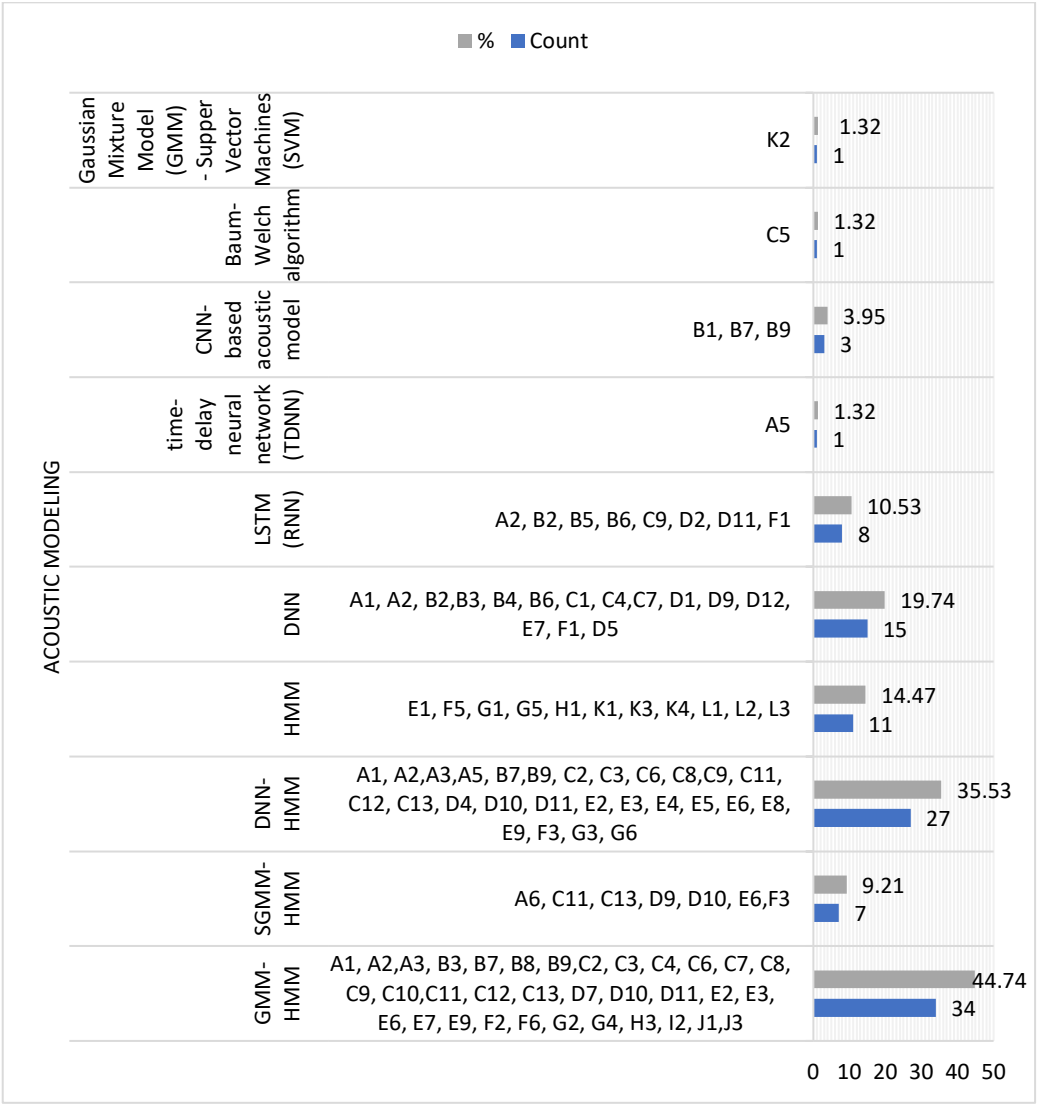

Figure 6. Acoustic models used by the researchers.

4.4. RQ 4: What Type of Speech Datasets Were Used during the Training and Testing of the Models in the Papers Included in the SLR?

The speech dataset used for developing speech recognition plays an important role in obtaining a higher speech recognition rate. From the studies, it was found that different datasets were used by the researchers to test and train the speech recognition system. Datasets can be either private or public, and most of the datasets were publicly available on the web. Most of the researchers used multiple datasets for training and testing the ASR system. Table 5 gives information on the datasets used in the research papers. From the table, it is clear that the datasets WSJCAM0 and PF-STAR are the first choice of the researchers to train and test the children's ASR system. The WSJCAM0 dataset was used in 26 research papers and the PF-STAR dataset was also used in 26 research papers. In 13 research papers, researchers preferred the TIDIGIT dataset. Along with this finding, ChildIt, TIMIT, and the CMU Kid dataset were used in 7, 4, and 4 papers, respectively. The rest of the papers used the researchers' private or other datasets.

Table 5. Corpus details.

Corpus Name	Paper ID	Count	%
WSJCAM0	A2, A3, B2, B5, B6, B7, C2, C3, C6, C8, C9, C10, C11, C12, C13, D2, D4, D9, D10, E3, E6, E7, F5, F6, H3, K3	26	34.21
PF-STAR	A2, A3, B2, B5, B6, B7, C1, C2, C6, C8, C9, C10, C11, C13, D2, D4, D7, D9, D10, E3, E6, E7, E8, F5, F6, K3	26	34.21
TIDIGIT	B8, C3, C11, F5, G4, H2, H3, I2, J1, K1, K3, K4, L3	13	17.11
ChildIt	E4, E5, F3, G3, G6, J3, L1	7	9.21
TIMIT	E1, E9, F2, L2	4	5.26
CMU Kid	D5, D11, E9, H3	4	5.26
Other	A1, A4, A5, A6, B1, B9, C4, C5, C7, D1, D3, D6, D8, D12, E2, F1, F4, G1, G2, G5, H1, I1, J2, K2	24	32.89

4.5. RQ 5: Which Toolkits Are Used by the Authors for Speech Recognition?

This section presents the speech recognition toolkits used by the researchers to recognize the speech signals. Figure 7 and Appendix B gives detailed information on the toolkits used. From the studies, it was found that almost half of the papers (48.68%) used the Kaldi toolkit for performing the experiments and was the first choice of the researchers. In certain papers (21.05%), the authors used the HTK toolkit for developing the ASR system, while CMU Sphinx was used in 2.63% of papers. In 27.63% of studies, the authors have not mentioned the name of the ASR toolkit used for performing the experiments.

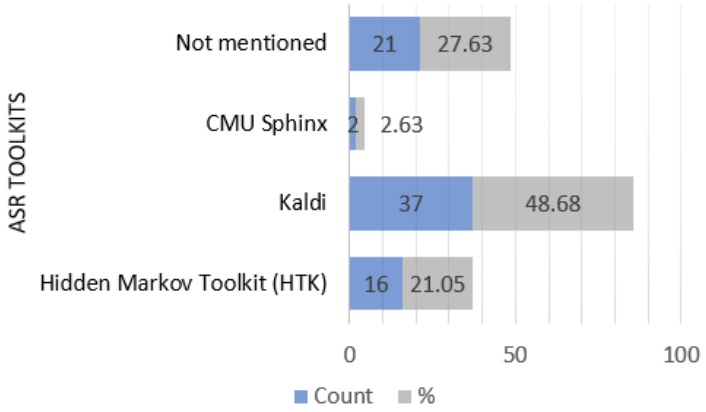

Figure 7. Toolkits used for speech recognition.

4.6. RQ 6: What Are the Different Natural Languages Used to Develop the ASR System?

In this section, we discuss the languages used by researchers to build a Children's ASR system. Figure 8 and Appendix C present information concerning the various languages used in the studies. In 63% out of 76 papers included in the systematic review, researchers used a dataset built using the English language to train and test the ASR system. The Italian language was used in 9% of the papers. In 5% of the papers, the Mandarin language is used by the researchers. The authors of the studies also developed children's speech recognition systems for the Malay, Russian, Japanese, Punjabi, German, and Portuguese languages. In 5% of the papers, the authors did not mention the language used, while 4% used multiple languages for training and testing the system.

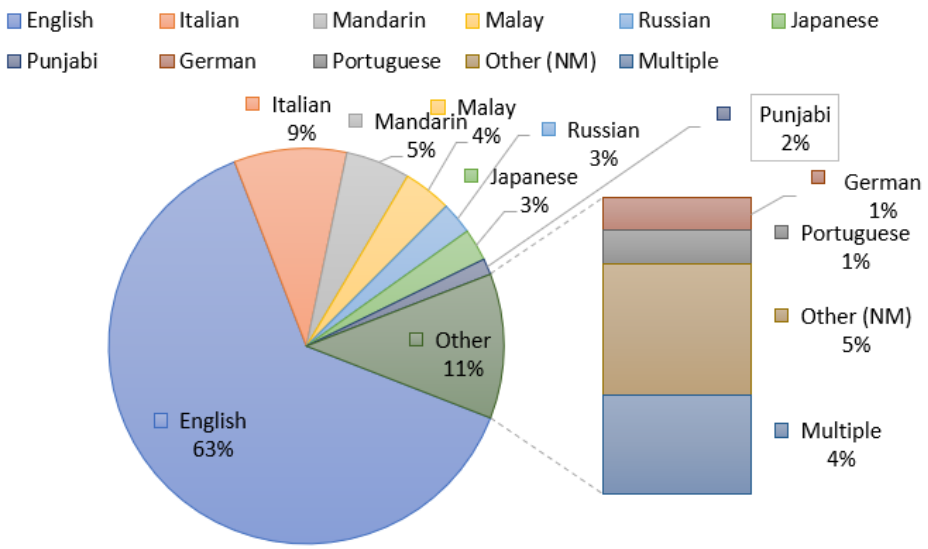

Figure 8. Languages used in the studies.

4.7. RQ 7: What Type of Environmental Conditions Were Used by Researchers to Develop the Speech Recognition System?

Environmental conditions play an important role in the development of the speech corpus for building an ASR system. Different environmental conditions were used by researchers to develop the speech corpus for training and testing the system, as shown in Figure 9. Neutral, noisy, and noise-free environmental conditions were used in the studies. In total, 86.8% of the papers used a neutral environment, while 9.2% used noise-free speech datasets for training and testing the system, and only 3.9% used noisy environmental conditions.

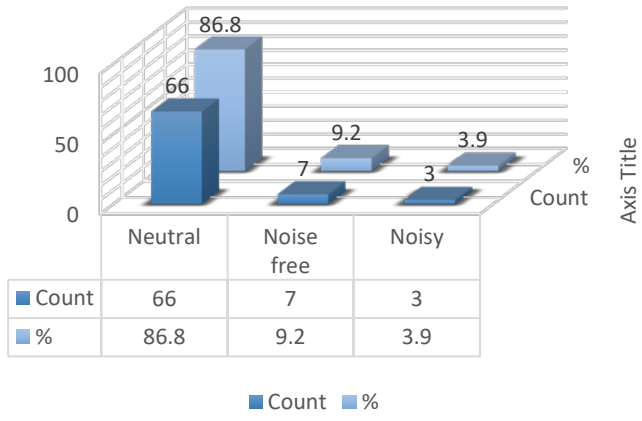

Figure 9. Environmental conditions used in the studies.

4.8. RQ 8: Which Are the Evaluation Methods Used to Evaluate the Overall Performance?

The authors used a variety of evaluation methodologies to assess the overall efficiency and performance of the proposed system. According to the research, the majority of researchers used the word error rate (WER) to evaluate the performance of the produced system. From Table 6, it is clear that 73.7% of the papers used WER as an evaluation method. The accuracy of the system was calculated in 7.9% of the papers to measure the performance of the system. The phone error rate (PER) is another evaluation method used in 6.6% of the studies. The rest of the papers used the sentence error rate, unweighted average recall (UAR), character error rate (CER), sentence correctness, and word correctness evaluation methods.

Table 6. Evaluation methods used.

Evaluation Method	Paper ID	Count	%
Word Error Rate (WER)	A1, A2, A3, A5, A6, B2, B3, B4, B5, B6, B7, B9, C1, C2, C3, C4,C6, C7, C8, C9, C10, C11, C12, C13, D2, D4, D5, D7, D9, D10, D11, D12, E2, E3, E5, E6, E7, E8, E9, F1, F3 F5, F6, G4, G5, H1, H2, H3, I2, J1, J2, K1, K3, K4, L2, L3	56	73.7
Sentence Error Rates	A1, B8	2	2.6
Unweighted Average Recall (UAR)	D6	1	1.3
Character Error Rate (CER)	B1	1	1.3
Accuracy Rate	C5, D6, E1, F2, I1, K2	6	7.9
Phone Error Rate	E4, G3, G6, J3, L1	5	6.6
Sentence Correctness	G1	1	1.3
Word Correctness	G1, K2	2	2.6

5. Discussion and Conclusions

In this paper, we have presented a systematic literature review of children's speech recognition systems studied from 2009 to 2020. The data and information gathered from studies on feature extraction, auditory modeling, datasets, different languages, and surroundings can be used to construct an ASR system for children. This also benefits researchers in conducting new studies to improve the recognition of children's speech. Additionally, data acquired from research publications helps in determining the research area trends and research gaps in this field. From the studies, it was found that research into the recognition of children's speech and its variations are in a very small group of studies. Recognition accuracy was also lagging compared to adult speech recognition.

After conducting the SLR it was found that most of the papers (62%) were published by the conferences, while Interspeech published 36.17% of the conference research papers. In the case of journals, the majority of the research papers (22.2%) were published by the Computer Speech & Language journal, while WOCCI is one of the popular workshops where 45.45% of the workshop papers were published.

In terms of extracting acoustic features from speech signals, the MFCC feature extraction technique is the one most used in 80% of the research conducted by researchers. Researchers prefer hybrid acoustic models instead of using standalone models for classification. The GMM-HMM hybrid model was used in 44.74% of the studies and the DNN-HMM model was used in 35.53% of the studies. Most of the work (63.16%) was conducted for the English language and far less work was conducted for other languages, namely, Italian, Mandarin, Malay, Punjabi, and many more. It was also found that the Kaldi toolkit is the most popular speech recognition toolkit and is the first choice of researchers when developing ASR systems. In the past five years, work conducted in the field of children's speech recognition has rapidly increased; the highest number of papers were published in 2018. This systematic review introduces the trends in the field of children's ASR for those researchers who have very little information about the topic, especially for those who are interested in children's speech recognition programs.

Author Contributions: Conceptualization, V.B., H.H. and V.K.; methodology, V.B., M.T.B.O. and Y.B.; software, V.B., M.B, B.S.G. and A.U.R.; validation, V.B., V.K., M.S. and H.H.; formal analysis, V.B.; investigation, V.K., M.S. and H.H.; resources, V.B., M.T.B.O. and Y.B.; data curation, V.B., M.B, B.S.G. and A.U.R.; writing—original draft preparation, V.B.; writing—review and editing, V.B. and A.U.R.; visualization, V.B.; supervision, V.K and H.H.; project administration, V.B., M.S. and M.B.; funding acquisition, M.T.B.O. and H.H. All authors have read and agreed to the published version of the manuscript.

Funding: The researchers would like to thank the Deanship of Scientific Research, Qassim University, for funding the publication of this project.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

Abbreviations

ASR	Automatic Speech Recognition
SR	Speech Recognition
SLR	Systematic Literature Review
STT	Speech to Text
HTK	Hidden Markov Toolkit
LVCSR	Large Vocabulary Continuous Speech Recognition
VTLN	Vocal Tract Length Normalization
MAP	Maximum a Posteriori
ML	Maximum-Likelihood
MLLR	Maximum Likelihood Linear Regression
cMLLR	Constrained MLLR
fMLLR	Feature Space MLLR
SAT	Speaker Adaptive Training
IVAs	Intelligent Virtual Assistants
RQ	Research Question
MFCC	Mel-Frequency Cepstral Coefficients
DNN	Deep Neural Network
WER	Word Error Rate
LDA	Linear Discriminant Analysis
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
SVM	Supper Vector Machines
CNN	Convolutional Neural Network
PER	Phone Error Rate
CER	Character Error Rate
UAR	Unweighted Average Recall
TDNN	Time-Delay Neural Network

Appendix A

Table A1. Assigned Paper IDs.

Paper ID	Reference Number	Paper ID	Reference Number	Paper ID	Reference Number	Paper ID	Reference Number
A1	[23]	A2	[24]	A3	[25]	A4	[26]
A5	[27]	A6	[21]	B1	[28]	B2	[29]
B3	[30]	B4	[31]	B5	[32]	B6	[33]
B7	[34]	B8	[35]	B9	[36]	C1	[37]
C2	[38]	C3	[39]	C4	[40]	C5	[41]
C6	[42]	C7	[43]	C8	[44]	C9	[45]
C10	[46]	C11	[47]	C12	[48]	C13	[49]

Table A1. *Cont.*

Paper ID	Reference Number	Paper ID	Reference Number	Paper ID	Reference Number	Paper ID	Reference Number
D1	[50]	D2	[51]	D3	[52]	D4	[53]
D5	[54]	D6	[55]	D7	[56]	D8	[57]
D9	[58]	D10	[59]	D11	[60]	D12	[61]
E1	[62]	E2	[63]	E3	[64]	E4	[65]
E5	[66]	E6	[67]	E7	[68]	E8	[69]
E9	[70]	F1	[71]	F2	[72]	F3	[73]
F4	[74]	F5	[75]	F6	[76]	G1	[77]
G2	[78]	G3	[79]	G4	[80]	G5	[81]
G6	[82]	H1	[83]	H2	[84]	H3	[85]
I1	[86]	I2	[87]	J1	[88]	J2	[89]
J3	[90]	K1	[91]	K2	[92]	K3	[93]
K4	[94]	L1	[95]	L2	[96]	L3	[97]

Appendix B

Table A2. SR Toolkits used in the studies.

Speech Recognition Toolkit	Paper ID
Hidden Markov Toolkit (HTK)	C10, E1, F2, F5, F6, G1, G4, H2, H3, I2, J1, K1, K3, K4, L2, L3
Kaldi	A1, A2, A5, A6, B2, B3, B4, B5, B6, B7, B8, B9, C1, C2, C3, C6, C7, C8, C9, C11, C12, C13, D2, D4, D5, D7, D9, D10, D11, E2, E3, E5, E6, E7, E8, F3, G5
CMU Sphinx	G3, J3
Not mentioned	A3, A4, B1, C4, C5, D1, D3, D6, D8, D12, E4, E9, F1, F4, G2, G6, H1, I1, J2, K2

Appendix C

Table A3. Natural languages used in the studies.

Language	Paper ID
English	A1, A2, A3, A4, A5, B2, B5, B6, B7, B8, B9, C2, C3, C6, C8, C9, C10, C11, C12, C13, D2, D4, D7. D9, D10, D11, E2, E3, E6, E7, E8, E9, F2, F3, F5, F6, G2, G4, G5, H2, H3, I2, J1, K1, K3, K4, L2, L3
Italian	E4, E5, F3, G3, G6, J3, L1
Mandarin	B3, B4, D1, D8
Malay	E1, G1, I1
Russian	A4, D6
Japanese	B1, C5
Punjabi	A6
German	K2
Portuguese	H1
Other (NM)	D3, D5, F4, J2
Multiple	C1, C4, D12

References

1. Ali, A.; Gravino, C. A systematic literature review of software effort prediction using machine learning methods. *J. Softw. Evol. Process* **2019**, *31*, e2211. [CrossRef]
2. De Lima, T.A.; Speech, C. A Survey on Automatic Speech Recognition Systems for Portuguese Language and its Variations. *Comput. Speech Lang.* **2019**, *62*, 101055. [CrossRef]
3. Claus, F.; Rosales, H.G.; Petrick, R.; Hain, H. A Survey about Databases of Children's Speech a Survey about Databases of Children's Speech Dresden University of Technology, Chair for System Theory and Speech Technology. *INTERSPEECH*. 2015, pp. 2410–2414. Available online: https://www.isca-speech.org/archive_v0/archive_papers/interspeech_2013/i13_2410.pdf (accessed on 15 March 2021).
4. HTK Speech Recognition Toolkit. Available online: http://htk.eng.cam.ac.uk/ (accessed on 2 September 2020).
5. Overview of the CMUSphinx Toolkit. Available online: https://cmusphinx.github.io/wiki/tutorialoverview/ (accessed on 2 September 2020).
6. Povey, D.; Ghoshal, A.; Boulianne, G. The Kaldi Speech Recognition Toolkit. *IEEE Signal Process. Soc.* **2011**, 1–4. Available online: http://kaldi.sf.net/ (accessed on 19 July 2020).
7. Open-Source Large Vocabulary CSR Engine Julius. Available online: http://julius.osdn.jp/en_index.php (accessed on 2 September 2020).
8. Sunil, Y.; Prasanna, S.R.M.; Sinha, R. Children's Speech Recognition under Mismatched Condition: A Review. *IETE J. Educ.* **2016**, *57*, 96–108. [CrossRef]
9. Taniya; Bhardwaj, V.; Kadyan, V. Deep Neural Network Trained Punjabi Children Speech Recognition System Using Kaldi Toolkit. In Proceedings of the 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 30–31 October 2020; pp. 374–378.
10. Claus, F.; Rosales, H.G.; Petrick, R.; Hain, H. A Survey about ASR for Children. *ISCA Archive*. 2013, pp. 26–30. Available online: https://www.isca-speech.org/archive_v0/slate_2013/papers/sl13_026.pdf (accessed on 5 July 2021).
11. Kathania, H.K.; Kadiri, S.R.; Alku, P.; Kurimo, M. Spectral modification for recognition of children's speech undermismatched conditions. In Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa); Linköping University Electronic Press: Linköping, Sweden, 2021; pp. 94–100. Available online: https://aclanthology.org/2021.nodalida-main.10 (accessed on 5 September 2021).
12. Madhavi, M.C.; Patil, H.A. Vocal Tract Length Normalization using a Gaussian mixture model framework for query-by-example spoken term detection. *Comput. Speech Lang.* **2019**, *58*, 175–202. [CrossRef]
13. Kathania, H.K.; Kadiri, S.R.; Alku, P.; Kurimo, M. A formant modification method for improved ASR of children's speech. *Speech Commun.* **2021**, *136*, 98–106. [CrossRef]
14. Tsao, Y.; Lai, Y.H. Generalized maximum a posteriori spectral amplitude estimation for speech enhancement. *Speech Commun.* **2016**, *76*, 112–126. [CrossRef]
15. Bhardwaj, V.; Kukreja, V. Effect of pitch enhancement in Punjabi children's speech recognition system under disparate acoustic conditions. *Appl. Acoust.* **2021**, *177*, 107918. [CrossRef]
16. Bhardwaj, V.; Kukreja, V.; Singh, A. Usage of Prosody Modification and Acoustic Adaptation for Robust Automatic Speech Recognition (ASR) System. *Rev. d'Intell. Artif.* **2021**, *35*, 235–242. [CrossRef]
17. Takaki, S.; Kim, S.; Yamagishi, J. Speaker Adaptation of Various Components in Deep Neural Network based Speech Synthesis. *Speech Synthesis Workshop*. 2016, pp. 153–159. Available online: https://206.189.82.22/archive_v0/SSW_2016/pdfs/ssw9_PS2-5_Takaki.pdf (accessed on 15 April 2021).
18. Kathania, H.K.; Kadiri, S.R.; Alku, P.; Kurimo, M. Using data augmentation and time-scale modification to improve asr of children's speech in noisy environments. *Appl. Sci.* **2021**, *11*, 8420. [CrossRef]
19. Kaur, H.; Bhardwaj, V.; Kadyan, V. Punjabi Children Speech Recognition System under Mismatch Conditions Using Discriminative Techniques. In *Innovations in Computer Science and Engineering*; Springer: Singapore, 2021; pp. 195–202.
20. Klejch, O.; Fainberg, J.; Bell, P.; Renals, S. Speaker Adaptive Training Using Model Agnostic Meta-Learning. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Sentosa, Singapore, 14–18 December 2019; pp. 881–888.
21. Bhardwaj, V.; Bala, S.; Kadyan, V.; Kukreja, V. Development of Robust Automatic Speech Recognition System for Children's using Kaldi Toolkit. In Proceedings of the Second International Conference on Inventive Research in Computing Applications (ICIRCA-2020), Coimbatore, India, 15–17 July 2020; pp. 10–13. [CrossRef]
22. Bala, S.; Kadyan, V.; Bhardwaj, V. Bottleneck Feature Extraction in Punjabi Adult Speech Recognition System. In *Innovations in Computer Science and Engineering*; Springer: Singapore, 2021; pp. 493–501.
23. Shivakumar, P.G.; Georgiou, P. Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations. *Comput. Speech Lang.* **2020**, *63*, 101077. [CrossRef]
24. Shahnawazuddin, S.; Bandarupalli, T.S.; Chakravarthy, R. Improving Automatic Speech Recognition by Classifying Adult and Child Speakers into Separate Groups using Speech Rate Rhythmicity Parameter. In Proceedings of the International Conference on Signal Processing and Communications (SPCOM), Bangalore, India, 28 August 2020; pp. 1–5. [CrossRef]

25. Kathania, H.K.; Kadiri, S.R.; Alku, P.; Kurimo, M. Study of formant modification for children ASR. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual Barcelona, 4–8 May 2020; pp. 7424–7428. [CrossRef]
26. Riekhakaynen, E.I. Corpora of Russian Spontaneous Speech as a Tool for Modelling Natural Speech Production and Recognition. In Proceedings of the Annual Computing and Communication Workshop and Conference, CCWC 2020, Las Vegas, NV, USA, 6–8 January 2020; pp. 406–411. [CrossRef]
27. Kumar, M.; Kim, S.H.; Lord, C.; Lyon, T.D.; Narayanan, S. Leveraging Linguistic Context in Dyadic Interactions to Improve Automatic Speech Recognition for Children. *Comput. Speech Lang.* **2020**, *63*, 101101. [CrossRef]
28. Nagano, T.; Fukuda, T.; Suzuki, M.; Kurata, G. Data Augmentation Based on Vowel Stretch for Improving Children's Speech Recognition. In Proceedings of the Automatic Speech Recognition and Understanding Workshop, ASRU, Singapore, 14–18 December 2019; pp. 502–508. [CrossRef]
29. Shahnawazuddin, S.; Adiga, N.; Kathania, H.K.; Sai, B.T. Creating speaker independent ASR system through prosody modification based data augmentation. *Pattern Recognit. Lett.* **2020**, *131*, 213–218. [CrossRef]
30. Sheng, P.; Yang, Z.; Qian, Y. GANs for Children: A Generative Data Augmentation Strategy for Children Speech Recognition. In Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Sentosa, Singapore, 14–18 December 2019; pp. 129–135.
31. Li, C.; Qian, Y. Prosody usage optimization for children speech recognition with zero resource children speech. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Graz, Austria, 15–19 September 2019; Volume 2019, pp. 3446–3450. [CrossRef]
32. Yadav, I.C.; Pradhan, G. Significance of Pitch-Based Spectral Normalization for Children's Speech Recognition. *IEEE Signal Process. Lett.* **2019**, *26*, 1822–1826. [CrossRef]
33. Yadav, I.C.; Shahnawazuddin, S.; Pradhan, G. Addressing noise and pitch sensitivity of speech recognition system through variational mode decomposition based spectral smoothing. *Digit. Signal Process. Rev. J.* **2019**, *86*, 55–64. [CrossRef]
34. Dubagunta, S.P.; Kabil, S.H.; Doss, M.M. Improving Children Speech Recognition through Feature Learning from Raw Speech Signal. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing-ICASSP, Brighton, UK, 12–17 May 2019; pp. 5736–5740. [CrossRef]
35. Naing, H.M.S.; Miyanaga, Y.; Hidayat, R.; Winduratna, B. Filterbank Analysis of MFCC Feature Extraction in Robust Children Speech Recognition. In Proceedings of the International Symposium on Multimedia and Communication Technology, ISMAC, Quezon City, Philippines, 19–21 August 2019; pp. 1–6. [CrossRef]
36. Rehman, A.U.; Naqvi, R.A.; Rehman, A.; Paul, A.; Sadiq, M.T.; Hussain, D. A Trustworthy SIoT Aware Mechanism as an Enabler for Citizen Services in Smart Cities. *Electronics* **2020**, *9*, 918. [CrossRef]
37. Matassoni, M.; Gretter, R.; Falavigna, D.; Giuliani, D. Non-Native Children Speech Recognition Through Transfer Learning. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Calgary, AB, Canada, 15–20 April 2018; pp. 6229–6233. [CrossRef]
38. Kathania, H.K.; Shahnawazuddin, S.; Ahmad, W.; Adiga, N.; Jana, S.K.; Samaddar, A.B. Improving children's speech recognition through time scale modification based speaking rate adaptation. In Proceedings of the International Conference on Signal Processing and Communications, Bangalore, India, 16–19 July 2018; pp. 257–261. [CrossRef]
39. Kathania, H.K.; Shahnawazuddin, S.; Adiga, N.; Ahmad, W. Role of Prosodic Features on Children's Speech Recognition. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Calgary, AB, Canada, 15–20 April 2018; pp. 5519–5523. [CrossRef]
40. Sabu, K.; Rao, P. Automatic assessment of children's oral reading using speech recognition and prosody modeling. *CSI Trans. ICT* **2018**, *6*, 221–225. [CrossRef]
41. Tian, Y.; Tang, J.; Jiang, X.; Tsutsui, H.; Miyanaga, Y. Accuracy on Children's Speech Recognition under Noisy Circumstances. In Proceedings of the Iternational Symposium on Communication and Information Technology, Bangkok, Thailand, 26–29 September 2018; pp. 101–104. [CrossRef]
42. Shahnawazuddin, S.; Singh, C.; Kathania, H.K.; Ahmad, W.; Pradhan, G. An Experimental Study on the Significance of Variable Frame-Length and Overlap in the Context of Children's Speech Recognition. *Circuits Syst. Signal Process.* **2018**, *37*, 5540–5553. [CrossRef]
43. Watson, S.; Coy, A. JAMLIT: A Corpus of Jamaican Standard English for Automatic Speech Recognition of Children's Speech. In Proceedings of the Workshop on Spoken Language Technologies for Under-Resourced Languages, Gurugram, India, 29–31 August 2018; pp. 238–242. [CrossRef]
44. Shahnawazuddin, S.; Kathania, H.K.; Singh, C.; Ahmad, W.; Pradhan, G. Exploring the role of speaking-rate adaptation on children's speech recognition. In Proceedings of the International Conference on Signal Processing and Communications, Bangalore, India, 16–19 July 2018; pp. 21–25. [CrossRef]
45. Yadav, I.C.; Kumar, A.; Shahnawazuddin, S.; Pradhan, G. Non-uniform spectral smoothing for robust children's speech recognition. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Hyderabad, India, 2–6 September 2018; pp. 1601–1605. [CrossRef]
46. Shahnawazuddin, S.; Sinha, R. A Fast Adaptation Approach for Enhanced Automatic Recognition of Children's Speech with Mismatched Acoustic Models. *Circuits Syst. Signal Process.* **2018**, *37*, 1098–1115. [CrossRef]

47. Kathania, H.K.; Ahmad, W.; Shahnawazuddin, S.; Samaddar, A.B. Explicit Pitch Mapping for Improved Children's Speech Recognition. *Circuits Syst. Signal Process.* **2018**, *37*, 2021–2044. [CrossRef]
48. Shahnawazuddin, S.; Adiga, N.; Kathania, H.K.; Pradhan, G.; Sinha, R. Studying the role of pitch-adaptive spectral estimation and speaking-rate normalization in automatic speech recognition. *Digit. Signal Process. Rev. J.* **2018**, *79*, 142–151. [CrossRef]
49. Sinha, R.; Shahnawazuddin, S. Assessment of pitch-adaptive front-end signal processing for children's speech recognition. *Comput. Speech Lang.* **2018**, *48*, 103–121. [CrossRef]
50. Tong, R.; Chen, N.F.; Ma, B. Multi-task learning for mispronunciation detection on Singapore children's Mandarin speech. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Stockholm, Sweden, 20–24 August 2017; pp. 2193–2197. [CrossRef]
51. Ganji, S.; Sinha, R. Exploring recurrent neural network based acoustic and linguistic modeling for children's speech recognition. In Proceedings of the Annual International Conference, Proceedings/TENCON, Penang, Malaysia, 5–8 November 2017; pp. 2880–2884. [CrossRef]
52. Grieco-Calub, T.M.; Ward, K.M.; Brehm, L. Multitasking during degraded speech recognition in school-age children. *Trends Hear.* **2017**, *21*, 1–14. [CrossRef]
53. Shahnawazuddin, S.; Deepak, K.T.; Pradhan, G.; Sinha, R. Enhancing noise and pitch robustness of children's ASR. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 5225–5229.
54. Kumar, M.; Bone, D.; McWilliams, K.; Williams, S.; Lyon, T.D.; Narayanan, S. Multi-scale context adaptation for improving child automatic speech recognition in child-adult spoken interactions. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Stockholm, Sweden, 20–24 August 2017; pp. 2730–2734. [CrossRef]
55. Kaya, H.; Salah, A.A.; Karpov, A.; Frolova, O.; Grigorev, A.; Lyakso, E. Emotion, age, and gender classification in children's speech by humans and machines. *Comput. Speech Lang.* **2017**, *46*, 268–283. [CrossRef]
56. Alharbi, S.; Simons, A.J.H. Automatic recognition of children's read speech for stuttering application. In Proceedings of the International Workshop on Child Computer Interaction, Glasgow, UK, 13–17 November 2017; pp. 1–6.
57. Zhou, H.; Li, Y.; Liang, M.; Guan, C.Q.; Zhang, L.; Shu, H.; Zhang, Y. Mandarin-speaking children's speech recognition: Developmental changes in the influences of semantic context and F0 contours. *Front. Psychol.* **2017**, *8*, 1–7. [CrossRef]
58. Shahnawazuddin, S.; Sinha, R.; Pradhan, G. Pitch-Normalized Acoustic Features for Robust Children's Speech Recognition. *IEEE Signal Process. Lett.* **2017**, *24*, 1128–1132. [CrossRef]
59. Ahmad, W.; Shahnawazuddin, S.; Kathania, H.K.; Pradhan, G.; Samaddar, A.B. Improving children's speech recognition through explicit pitch scaling based on iterative spectrogram inversion. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Stockholm, Sweden, 20–24 August 2017; pp. 2391–2395. [CrossRef]
60. Qian, Y.; Evanini, K.; Wang, X.; Lee, C.M.; Mulholland, M. Bidirectional LSTM-RNN for improving automated assessment of non-native children's speech. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Stockholm, Sweden, 20–24 August 2017; pp. 1417–1421. [CrossRef]
61. Tong, R.; Wang, L.; Ma, B. Transfer learning for children's speech recognition. In Proceedings of the International Conference on Asian Language Processing, IALP, Singapore, 5–7 December 2017; pp. 36–39. [CrossRef]
62. Mustafa, M.B. A Two-Stage Adaptation towards Automatic Speech Recognition System for Malay-Speaking Children. *Int. J. Comput. Electr. Autom. Control Inf. Eng.* **2016**, *10*, 513–516.
63. Qian, Y.; Wang, X.; Evanini, K.; Suendermann-Oeft, D. Improving DNN-Based Automatic Recognition of Non-native Children Speech with Adult Speech. In Proceedings of the Workshop on Child Computer Interaction, San Francisco, CA, USA, 6–7 September 2016; pp. 40–44. [CrossRef]
64. Sinha, R.; Shahnawazuddin, S.; Karthik, P.S. Exploring the role of pitch-adaptive cepstral features in context of children's mismatched ASR. In Proceedings of the 2016 International Conference on Signal Processing and Communications (SPCOM), Bangalore, India, 12–15 June 2016.
65. Serizel, R.; Giuliani, D. Deep neural network adaptation for children's and adults' speech recognition. In Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014, Pisa, Italy, 9–11 December 2014.
66. Matassoni, M.; Falavigna, D.; Giuliani, D. DNN adaptation for recognition of children speech through automatic utterance selection. In Proceedings of the Workshop on Spoken Language Technology, SLT 2016-Proceedings, San Diego, CA, USA, 13–16 December 2016; pp. 644–651. [CrossRef]
67. Kathania, H.K.; Shahnawazuddin, S.; Pradhan, G.; Samaddar, A.B. Experiments on children's speech recognition under acoustically mismatched conditions. In Proceedings of the Annual International Conference-TENCON, Singapore, 22–25 November 2016; pp. 3014–3017. [CrossRef]
68. Fainberg, J.; Bell, P.; Lincoln, M.; Renals, S. Improving children's speech recognition through out-of-domain data augmentation. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, San Francisco, CA, USA, 8–12 September 2016; pp. 1598–1602. [CrossRef]
69. Shahnawazuddin, S.; Dey, A.; Sinha, R. Pitch-adaptive front-end features for robust children's ASR. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, San Francisco, CA, USA, 8–12 September 2016; pp. 3459–3463. [CrossRef]

70. Qian, M.; McLaughlin, I.; Quo, W.; Dai, L. Mismatched training data enhancement for automatic recognition of children's speech using DNN-HMM. In Proceedings of the 2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP), Tianjin, China, 17–20 October 2016. [CrossRef]
71. Liao, H.; Pundak, G.; Siohan, O.; Carroll, M.; Coccaro, N.; Jiang, Q.M.; Sainath, T.N.; Senior, A.; Beaufays, F.; Bacchiani, M. Large vocabulary automatic speech recognition for children. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Dresden, Germany, 6–10 September 2015; pp. 1611–1615.
72. Fringi, E.; Lehman, J.F.; Russell, M. Evidence of phonological processes in automatic recognition of children's speech. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Dresden, Germany, 6–10 September 2015; pp. 1621–1624.
73. Giuliani, D.; BabaAli, B. Large vocabulary children's speech recognition with DNN-HMM and SGMM acoustic modeling. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Dresden, Germany, 6–10 September 2015; Volume 2015, pp. 1635–1639.
74. Nittrouer, S.; Tarr, E.; Wucinich, T.; Moberly, A.C.; Lowenstein, J.H. Measuring the effects of spectral smearing and enhancement on speech recognition in noise for adults and children. *J. Acoust. Soc. Am.* **2015**, *137*, 2004. [CrossRef] [PubMed]
75. Ghai, S.; Sinha, R. Pitch adaptive MFCC features for improving children's mismatched ASR. *Int. J. Speech Technol.* **2015**, *18*, 489–503. [CrossRef]
76. Shahnawazuddin, S.; Kathania, H.K.; Sinha, R. Enhancing the recognition of children's speech on acoustically mismatched ASR system. In Proceedings of the TENCON 2015–2015 IEEE Region 10 Conference, Macao, China, 1–4 November 2015. [CrossRef]
77. Rahman, F.D.; Mohamed, N.; Mustafa, M.B.; Salim, S.S. Automatic speech recognition system for Malay speaking children: Automatic speech recognition system. In Proceedings of the International Senior Project Conference, ICT-ISPC, Nakhonpathom, Thailand, 26–27 March 2014; pp. 79–82. [CrossRef]
78. Gray, S.S.; Willett, D.; Lu, J.; Pinto, J.; Maergner, P.; Bodenstab, N. Child Automatic Speech Recognition for US English: Child Interaction with living-room-electronic-devices. In Proceedings of the 4th Workshop on Child Computer Interaction (WOCCI 2014), Singapore, 19 September 2014; pp. 21–26.
79. Cosi, P.; Nicolao, M.; Paci, G.; Sommavilla, G.; Tesser, F. Comparing Open Source ASR Toolkits on Italian Children Speech. In Proceedings of the Workshop on Child Computer Interaction (WOCCI 2014), Singapore, 19 September 2014; pp. 1–6.
80. Sunil, Y.; Sinha, R. Exploration of MFCC based ABWE for robust children's speech recognition under mismatched condition. In Proceedings of the International Conference on Signal Processing and Communications, SPCOM 2014, Bangalore, India, 22–25 July 2014; pp. 1–5. [CrossRef]
81. 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 Interaction (WOCCI), Singapore, 19 September 2014; pp. 15–19.
82. Serizel, R.; Giuliani, D. Vocal tract length normalisation approaches to DNN-based children's and adults' speech recognition. In Proceedings of the Workshop on Spoken Language Technology, South Lake Tahoe, NV, USA, 7–10 December 2014; pp. 135–140. [CrossRef]
83. Hämäläinen, A.; Pinto, F.M.; Rodrigues, S.; Júdice, A.; Silva, S.M.; Calado, A.; Dias, M.S. A Multimodal Educational Game for 3-10-Year-Old Children: Collecting and Automatically Recognising European Portuguese Children's Speech. *SLaTE* **2013**, 31–36. Available online: http://www.isca-speech.org/archive (accessed on 3 October 2021).
84. Kathania, H.K.; Ghai, S.; Sinha, R. Soft-weighting technique for robust children speech recognition under mismatched condition. In Proceedings of the Annual IEEE India Conference, INDICON, Mumbai, India, 13–15 December 2013; pp. 1–6. [CrossRef]
85. Sanand, D.R.; Svendsen, T. Synthetic speaker models using VTLN to improve the performance of children in mismatched speaker conditions for ASR. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Lyon, France, 25–29 August 2013; pp. 3361–3365.
86. Zourmand, A.; Nong, T.H. Vowel classification of children's speech using fundamental and formant frequencies. In Proceedings of the International Conference on Computational Intelligence, Modelling and Simulation, Kuantan, Malaysia, 25–27 September 2012; pp. 282–287. [CrossRef]
87. Sunil, Y.; Sinha, R. Exploration of class specific ABWE for robust children's ASR under mismatched condition. In Proceedings of the International Conference on Signal Processing and Communications, SPCOM, Bangalore, India, 22–25 July 2012; pp. 1–5. [CrossRef]
88. Ghai, S.; Sinha, R. A study on the effect of pitch on LPCC and PLPC features for children's ASR in comparison to MFCC. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Florence, Italy, 27–31 August 2011; pp. 2589–2592.
89. Moya, E.; Hernandez, M.; Pineda, L.; Meza, I. Speech recognition with limited resources for children and adult speakers. In Proceedings of the International Conference on Artificial Intelligence: Advances in Artificial Intelligence and Applications, Puebla, Mexico, 26 November–4 December 2011; pp. 57–62. [CrossRef]
90. Nicolao, M.; Cosi, P. Comparing SPHINX vs. SONIC Italian Children Speech Recognition Systems. In Proceedings of the Conference of the Italian Association of Speech, Florence, Italy, 27–31 August 2011; pp. 1–12. Available online: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Comparing+SPHINX+vs+.+SONIC+Italian+Children+Speech+Recognition+Systems#0 (accessed on 3 October 2021).

91. Ghai, S.; Sinha, R. Analyzing pitch robustness of PMVDR and MFCC features for children's speech recognition. In Proceedings of the International Conference on Signal Processing and Communications, SPCOM, Bangalore, India, 18–21 July 2010; pp. 1–5. [CrossRef]
92. Bocklet, T.; Maier, A.; Eysholdt, U.; Nöth, E. Improvement of a speech recognizer for standardized medical assessment of children's speech by integration of prior knowledge. In Proceedings of the Workshop on Spoken Language Technology, SLT, Berkeley, CA, USA, 12–15 December 2010; pp. 259–264. [CrossRef]
93. Ghai, S.; Sinha, R. Enhancing children's speech recognition under mismatched condition by explicit acoustic normalization. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Chiba, Japan, 26–30 September 2010; pp. 522–525.
94. Ghai, S.; Sinha, R. Exploring the effect of differences in the acoustic correlates of adults' and children's speech in the context of automatic speech recognition. *Eurasip J. Audio Speech Music Process.* **2010**, *2010*, 318785. [CrossRef]
95. Cosi, P. On the development of matched and mismatched Italian children's speech recognition systems. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Brighton, UK, 6–10 September 2009; pp. 540–543.
96. Sinha, R.; Ghai, S. On the use of pitch normalization for improving children's speech recognition. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Brighton, UK, 6–10 September 2009; pp. 568–571.
97. Ghai, S.; Sinha, R. Exploring the role of spectral smoothing in context of children's speech recognition. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Brighton, UK, 6–10 September 2009; pp. 1607–1610.