				1. Qualitative 2. Quantitative		**1. Sample size 2. Sex 3. Age 4. Typically developing children/children with disabilities**	**1. Novel synthesis system 2. Commercial synthesis system 3. Review**
	Authors	**Year**	**Aims/purpose**	**Design**	**Language**	**Study population and sample size**	**Method**
1	Begnum, M.; Flatebø Hoelseth, S.; Johnsen, B.; Hansen, F.	2012	a) Could this strategy be successful for producing a high-quality child synthesis based on available Norwegian HMM voices? b) Aimed at exploring its appropriateness and fit for a specific user group (comparing adult and child voice output)	1. Qualitative: Questionnaires, interviews (talking maps interviews) and observations	Norwegian	1. Three children ("primary" testers) with severe expressive communication impairments (was four, one excluded) to participate in the user tests, along with their parents, schools and teachers ("secondary" testers). 2. Male. 3. 6-13 years. 4. Children with severe expressive communication impairments	1. Novel: Prototype an artificial Norwegian child voice based on speaker adaptive adjustment of adult based HMM synthesis.
2	Cosi, P	2015	In this paper, the KALDI ASR engine adapted to Italian is described and the results obtained on some children speech ASR experiments are reported.	2 Quantitative. Experimental design	Italian	Italian FBK ChildIt Corpus was used. It has 10 hours of speech from 171 children.	1. Novel. The Kaldi toolkit. KALDI currently contains two parallel implementations for DNN training.
3	Cosi, P.	2009	In this work the recent advances in the development of the SONIC Italian children's speech recognition system is described. This work, completing a previous one developed in the past, was conducted with the specific goals of integrating the newly trained children's speech recognition models into the Italian version of the Colorado Literacy Tutor platform.	2. Quantitative. Experimental	Italian	ChildIt corpus. Consists of Italian read sentences collected from 171 children (86 male and 85 female) aged between 7 and 13.	1 Novel. VTLN and SMAPLR.
4	Cosi, P.; Nicolao, M.; Paci, G.; Sommavilla, G.; Tesser, F.	2014	This paper aims at making a useful comparison among some of the best open-source ASR systems, and of demonstrating their effectiveness on a difficult target task, such as the application of ASR systems on children speech.	2. Quantitative. Experimental	Italian	ChildIt corpus. Consists of Italian read sentences collected from 171 children (86 male and 85 female) aged between 7 and 13.	1 Novel. The SONIC, SPHINX and BAVIECA ASR systems have been considered in the tests because they have similar structure, and all the three systems use Hidden Markov Models (HMM).

5	Cui, X.; Alwan, A.	2006	In this paper, we first study, in the discrete frequency domain, the relationship between frequency warping in the frontend and corresponding transformations in the back-end. Three common feature extraction schemes are investigated and their transformation linearity in the back-end are discussed.	1 Quantitative. Experimental	English	TIDIGIT database: contains 11.4 hours of speech data from 326 speakers (111 men, 114 women, 50 boys and 51 girls) uttering one to seven digits long strings consisting of eleven different digits (0-9 and 'OH'). The range of the age of the adult speakers is from 17 years to 70 years while the children speakers belong to an age from 6 years to 15 years.	1 Novel. VTLN, MLLR and a novel technique where speaker specific F3 information is taken for alignment, is compared.
6	Drager, K.; Reichle, J.; Pinkoski, C.	2010	A scoping review of the literature addressing the intelligibility and listener comprehension of synthesized speech output with children.	1. Qualitative: Scoping review	Unknown	1. 10 studies investigating synthesized speech intelligibility and/or comprehension with **children as listeners** were systematically identified and coded.	3. Review. Ten studies were identified. Studies included in the review were those that (a) were published between 1990 and 2009; (b) included children age 2–21 years; (c) included data on synthesized speech stimuli consisting of words, sentences, or phrases, or conversational intelligibility or comprehension using speech synthesis software (studies addressing intelligibility at the level of isolated sounds were excluded); and (d) investigated speech synthesis software still in use in AAC applications (SGDs).
7	Drager, K..; Clark-Serpentine, E.; Johnson, K.; Roeser, J.	2006	The present study investigated the intelligibility of digitized and synthesized speech output in background noise for children 3-5 years old. The purpose was to determine whether there was a difference in the intelligibility of 3 types of speech output (digitized, DECTalk synthesized, and MacinTalk synthesized) in single words and sentences, presented within and out of context.	1. Qualitative 2. Quantitative: Mixed modal research design (speech intelligibility test). Participants were asked to repeat 20 words and 10 short sentences.	Unknown	1. Ninety (3 x 30). 2. Equal numbers of boys and girls. 3. 3;0 to5.11-year-olds were assigned to each of the 3 speech type conditions. 4. Typically developing children).	2. Commercial synthesis systems: digitized, DECTalk synthesized, and MacinTalk synthesized. The digitized speech was the digital recording of an 11-year-old female speaker. The MacinTalk was "junior". The DECTalk voice was "Kit the Kid".

8	Drager, K..; Finke, E..	2012	The purpose of the study was to answer the following questions: First, what is the intelligibility of single word messages on digitized speech using an AAC system, recorded by typically developing children ages 4, 5, 6, and 7? Second, is there a cut-off age at which typically developing children's speech is more intelligible on digitized speech using an AAC system?	2. Quantitative: Forty adults were asked to transcribe 120 words spoken by child speakers in one of the age groups, presented on an AAC device with a digitized speech output.	Unknown	*Child Speakers:* 1. 16 children. 2. 2 boys and 2 girls in each of the four age groups (4-, 5-, 6- and 7-year-olds) 3. Between 4-7 years old 4. Typically developing children. *Adult Listeners.* 1. A total of 40 adult listeners (10 per age group). 3. Between the ages of 18 and 26 years.	1. Novel: The current investigation examined the intelligibility of digitized speech recorded from typically developing child speakers and reproduced on a AAC system with digitized speech.
9	Fainberg, J.; Bell, P.; Lincoln, M.; and Renals, S.	2016	In this work they 1) aim to build a strong baseline reporting up to 15.5% absolute improvements using Deep Neural Network (DNN) acoustic models and a stronger language model. 2) They then investigate un-modified augmentation, showing that contrary to the literature, DNN models may be able to use of out-of-domain data effectively. 3) Finally, they compare standard SFM with a novel use of SFM on out of- domain data	2 Quantitative. Experimental.	British English	English PF-STAR corpus: Augmentation with A) WSJCAM0 Corpus: 140 speakers between 18-28 years old. Approximately 81 hours of speech. B) ABI (Accents of the British Isles): 280 speakers, evenly distributed across gender and 14 accent groups. Randomly split, yielding about 16 hours of training data.	1. Novel. In this paper we explore data augmentation for children's speech recognition using stochastic feature mapping (SFM) to transform out-of-domain adult data for both GMM-based and DNN-based acoustic models. The Kaldi toolkit was used.
10	Fringi, E.; Lehman, J.; Russell, M.	2015	This study intended to look into the findings of speech development research and investigate whether any of the recorded phonological processes are reflected in the performance of a baseline recognizer through systematic error patterns.	1 Quantitative. Experimental	US English	The data used was collected from 60 students (10 five-year-olds, 16 six-year-olds, 14 seven-year-olds, 13 eight-year-olds and 17 nine-year-olds) from the state of Pennsylvania, U.S, ranging from pre-kindergardeners to third graders. The resulting phone confusion matrices are compared with one obtained for adult speech from a phone recognition experiment on the TIMIT corpus	1 Novel. Two tied-state triphone HMM-based ASR systems were developed, based on the CMU phone set, using the HTK toolkit.

11	Gerosa, M.; Giuliani, D.; Brugnara, F.	2007	This paper presents several acoustic analyses carried out on children's read speech. These analyses aimed at achieving a better understanding of spectral and temporal changes in speech produced by children of various ages in view of the development of automatic speech recognition applications.	2. Quantitative. Experimental	Italian	Several speech corpora were used in this work: three consisting of children's speech and two consisting of adult speech. 1) The ChildIt corpus: The ChildIt corpus is an Italian, task-independent, speech corpus that consists of clean read speech from children aged from 7 to 13 years, with a mean age of 10 years. 2) The SpontIt corpus is a task independent Italian speech database that consists of clean spontaneous speech from 21 children aged between 8 and 12, with a mean age of 10 years. 3) The CID corpus: an American English, task-independent, speech database that consists of read speech from 436 children aged from 5 to 18 and from 56 adult speakers. 1) The APASCI speech corpus: APASCI was developed at ITC-irst and consists of speech data collected from 176 adult speakers, gender balanced. Only a portion of APASCI corpus, consisting of speech from 124 speakers for a total of 5 h:38 m, was used in this work. 2) The IBN speech corpus is used for training the automatic broadcast news (BN) transcription system developedat ITC-irst for the Italian language and consists of about 5 h:48 m of speech.	1 Novel. HMM. Speaker adaptive acoustic modeling was investigated through the use of the VTLN, CMLSN and SAT methods.
12	Gerosa, M.; Giuliani, D.; Brugnara, F.	2009	In this paper, age-independent acoustic modeling is investigated in the context of large vocabulary speech recognition. Exploiting a small amount (9 hours) of children's speech and a more significant amount (57 hours) of adult speech, age-independent acoustic models are trained using several methods for speaker adaptive acoustic modeling.	1 Quantitative. Experimental	Italian	Five Italian speech corpora were used in this work. Three of these corpora consist of children's speech: the ChildIt corpus, the SpontIt corpus and the Tgr-child corpus. The other two corpora consist of adult speech: the IBN corpus and the Tgr-adult corpus.	1 Novel. The investigated solutions include cepstral mean and variance normalization, speaker adaptive acoustic modeling techniques, for acoustic models based on continuous density hidden Markov models (HMMs), and unsupervised speaker adaptation.
13	Gerosa, M.; Giuliani, D.; Narayanan, S; Potamianos, A.	2009	In this paper, the authors review: (1) the acoustic and linguistic properties of children's speech for both read and spontaneous speech, and (2) the performance of automatic speech recognition for children with application to spoken dialogue and multimodal dialogue system design.	2. Review.	Various	Discussed many child speech corpora.	3. Review. Discussed various systems.

14	Ghai, S.; Sinha, R.	2015	A pitch normalization algorithm is proposed for addressing the pitch mismatch between adults' and children's speech for children's automatic speech recognition (ASR).	2. Quantitative. Experimental	English	For experiments on a connected-digit recognition task, the TIDIGITS speech corpus is used for both adult and child speech data. For the continuous speech recognition task, the WSJCAM0 Cambridge Read News corpus is used for adult speech data and the PFSTAR British English corpus is used for child speech data.	1 Novel. Motivated by the appearance of pitch-dependent distortions in the smoothed melspectral envelope for high pitched children's speech, the algorithm modifies the mel filterbank during MFCC feature extraction to improve ASR performance.
15	Ghai, S.; Sinha, R.	2010	In this work, the researchers analyse the robustness of PMVDR to pitch variations across speech signals in comparison to MFCC for the children's speech recognition under mismatched condition.	2. Quantitative. Experimental	English	The TIDIGITS database contains 11.4 hours of speech data from 326 speakers (111 men, 114 women, 50 boys and 51 girls) uttering one to seven digits long strings consisting of eleven different digits 0-9 and 'OH'). The range of the age of the adult speakers is from 17 years to 70 years while the children speakers belong to an age from 6 years to 15 years.	1 Novel. The PMVDR feature to pitch variations across speech signals is compared to the commonly used MFCC feature.

16	Ghai, S.; Sinha, R.	2010	The efficacy of the researcher's recently proposed explicit acoustic (pitch and speaking rate) normalization in combination with the existing normalization/adaptation techniques is explored for mismatched children's speech recognition.	2. Quantitative. Experimental	English	The TIDIGITS database contains 11.4 hours of speech data from 326 speakers (111 men, 114 women, 50 boys and 51 girls) uttering one to seven digits long strings consisting of eleven different digits (0-9 and 'OH'). The range of the age of the adult speakers is from 17 years to 70 years while the children speakers belong to an age from 6 years to 15 years.	1 Novel. In this work, to study the effect of normalization of pitch and speaking rate, the average pitch and the duration of all children's speech test set signals are transformed to different values using the pitch synchronous time scaling (PSTS) method.
17	Ghai, S.; Sinha, R.	2010	This paper explores the independent effect of all of the acoustic sources of mismatch between adults' and children's speech reported in literature, that is, the pitch, the speaking rate, the formant frequencies, and the glottal parameters: OQ, RQ, and SQ on the recognition performance on a linguistically neutral task. Among these different acoustic sources of mismatch, the independent effects of the pitch and the glottal parameters on ASR have not been reported so far. The study is done on a limited vocabulary task (i.e., digit recognition) where the linguistic differences would be minimal.	2. Quantitative. Experimental	English	The TIDIGITS database contains 11.4 hours of speech data from 326 speakers (111 men, 114 women, 50 boys and 51 girls) uttering one to seven digits long strings consisting of eleven different digits (0-9 and 'OH'). The range of the age of the adult speakers is from 17 years to 70 years while the children speakers belong to an age from 6 years to 15 years.	1 Novel. In this work, the pitch, the signal duration (for modifying the speaking rate), and the glottal parameters, namely, the OQ, the RQ, and the SQ of the speech signals, are modified using a recently proposed pitch-synchronous time-scaling (PSTS) method. For addressing the mismatch in the formant frequencies between adults' and children's speech, the commonly used frequency warping is employed. For warping the frequency axis of the utterances during computation of the mel frequency cepstral coefficients (MFCCs) feature, the piece-wise linear frequency warping of filterbank, as supported in the Hidden Markov toolkit (HTK), has been used.

18	Ghai, S.; Sinha, R.	2009	In this paper, the researchers study the role of spectral smoothing in context of children's speech recognition. The spectral smoothing has been affected in the feature domain by two approaches; modification of bandwidth of the filters in the filterbank and cepstral truncation.	2. Quantitative. Experimental	English	The TIDIGITS database contains 11.4 hours of speech data from 326 speakers (111 men, 114 women, 50 boys and 51 girls) uttering one to seven digits long strings consisting of eleven different digits (0-9 and 'OH'). The range of the age of the adult speakers is from 17 years to 70 years while the children speakers belong to an age from 6 years to 15 years.	1 Novel. in this paper the researchers have experimented with the role of spectral smoothing in context of children's speech recognition. Initially, the bandwidth of the lower order filters in the filterbank was modified so as to smooth out the pitch-dependent distortions observed in the low frequency region of the spectral envelope. In addition to this, the experiments with varying cepstral truncation have also been performed so as to reduce the influence of the increase in variances of the higher cepstral coefficients with increasing pitch on the recognition performance. This in turn smooths the corresponding spectra implicitly. Both methods have shown the effectiveness of spectral smoothing in improving children's speech recognition. Apart from this, the effect of smoothing has also been studied in conjunction with the widely used VTLN in context of children's speech recognition.
19	Giuliani, D and BabaAli, B	2015	In this paper they investigate children's speech recognition by comparing SGMM and DNN-HMM systems in a large vocabulary recognition task when only a small amount of training data is available.	2 Quantitative. Experimental.	Italian	This study is conducted on the ChildIt speech corpus. The ChildIt corpus is an Italian, task-independent, speech corpus that consists of clean read speech from children aged from 7 to 13 years (roughly 7 hours of speech)	1 Novel. In this paper, large vocabulary children's speech recognition is investigated by using the Deep Neural Network – Hidden Markov Model (DNN-HMM) hybrid and the Subspace Gaussian Mixture Model (SGMM) acoustic modeling approach. The Kaldi toolkit was used.
20	Giuliani, D.; Gerosa, M.; Brugnara, F.	2006	In this paper, speaker adaptive acoustic modeling is investigated by using a novel method for speaker normalization and a well-known vocal tract length normalization method.	2. Quantitative. Experimental	Italian	The ChildIt corpus: Consists of Italian read sentences collected from 171 children (86 male and 85 female) aged between 7 and 13.	1 Novel. With the novel normalization method, acoustic observations of training and testing speakers are mapped into a normalized acoustic space through speaker-specific transformations with the aim of reducing inter-speaker acoustic variability. For each speaker, an affine transformation is estimated with the goal of reducing the mismatch between the acoustic data of the speaker and a set of target hidden Markov models. This transformation is estimated through constrained maximum likelihood linear regression and then applied to map the acoustic observations of the speaker into the normalized acoustic space.

21	Govender, A.; De Wet, F.	2016	The aim of this study was to determine whether a relationship exists between objective and subjective voice evaluation measures with regards to the selection of training speakers for an average voice model used in child speaker-adaptive HMM- based synthesis.	2. Quantitative: Experimental design.	US English	Children: 1. Four speakers were selected from the CMU-ARCTIC database. 2. English male child target speaker (using 100 sentences as adaptation data). [Four speaker dependent voices were built, one for each of the four training speakers. Each of these voices were then adapted to the target speaker.] Adult listeners: 1. 19 adult listeners native to South Africa.	1. Novel. Speaker adaptation in conjunction with Hidden Markov Model (HMM)-based synthesis.
22	Govender, A.; Nouhou, B.; De Wet, F.	2015	This study investigates the possibility of using data that has been recorded for ASR development to build a child voice for text-to-speech conversion in an HMM synthesis framework. The child voices generated using ASR data were compared with a voice built with data recorded specifically for TTS in terms of intelligibility and naturalness.	2. Quantitative: Experimental design. Evaluation 1: The evaluation consisted of an intelligibility test: A word error rate was calculated for each transcription and the voices were ranked. Another test was performed by just listening to the voices that had a WER less than 50%. In this manner, five voices were selected and ranked in terms of naturalness. Evaluation 2: For this round of evaluations, a formal listening test was conducted using a web-based evaluation test. The listening test consisted of two sections: a) Mean-Opinion-Score (MOS) test to evaluate natural- ness and b) transcription test to test intelligibility.	US English (no SA English child databases available)	1.One 2. boy child target speaker. The CMU kids corpus was used in this study. The corpus was recorded for the purpose of ASR development and comprises recordings of children in the 3. age group of 6 to 11 years old. 4. Typically developing. The **corpus** consists of 76 speakers in total of which 24 are male and 52 female.	1. Novel: HMM Model adaptation. The results indicated that a mixed average voice model (AVM) trained using one adult male (bdl) and one adult female (clb) speaker from the CMU- ARCTIC database was the best initial model for adaptation to a boy child target speaker.

23	Gray, S.; Willett, D.; Lu, J.; Pinto, J.; Maergner, P.; Bodenstab, N.	2014	In this study, the authors adapt multiple components of a state-of-the-art adult centric large vocabulary continuous speech recognition (LVCSR) system to build a child LVCSR system for living room electronic devices (LRED). Due to insufficient amounts of child-speech training data to build a child ASR system from scratch, they aimed to adapt their acoustic and language models from the adult-centric ASR to child ASR using a small amount of child data.	1 Quantitative. Experimental	US English	The data used in this study include speech from both adults and children. However, the majority of this data is neither manually marked as child vs. adult speech nor is the age of the speaker known. They evaluate their child-adapted ASR models on 6.8 hours of manually transcribed, LRED domain, child's speech. This data came from two different data collection sources, where speakers had used a speech recognition system in living room settings.	1 Novel. A GMM based model made with several thousand hours of transcribed data. The ASR system uses MFCC and delta coefficients as speech feature input, and the acoustic model makes use of context-dependent tree-based clustered Gaussian mixture HMMs. Acoustic modelling adaptation techniques they explored include vocal tract length normalization (VTLN), Maximum Likelihood Linear Regression (MLLR) and Constrained Maximum Likelihood Linear Regression (CMLLR).
24	Hagen, A.; Pellom, B.; Hacioglu, K.	2009	This work focuses on generating children's HMM-based acoustic models for speech recognition from adult acoustic models. Therefore, for a new language, an effective synthetic children's acoustic model can be derived without the need of children's data.	2. Quantitative: Experimental design. There was an objective evaluation of the voices using equal error rate between systems.	US English and Spanish	For a new language where only adult data is available, 1. an adult 2. male and an adult female model is trained. For both languages sufficient male, female, and children's speech data was available (more than 20 hours) in order to train valid acoustic models and to have reference children's acoustic models available. For English test data they used a corpus of 1. 22 native speakers in the 3. age range of 5 to 14. For Spanish test data the corpus is comprised of 1. 19 speakers in the 3. age range of 8 to 13 years.	1. Novel. Given this transitive chain of argumentation, the transformation from a male to a female acoustic model can be estimated for a language and applied (at a certain adjustable degree) to the female model. This process results in a synthetic children's speech model designed on the basis of the female model. Parameters that are subject to transformation in this process are the mean vectors of the HMM states. Therefore, for a new language an effective synthetic children's acoustic model can be derived without the need of children's data.
25	Hasija, T.; Kadyan, V.; Guleria, K.	2021	To overcome limited data conditions, in this paper, two corpora of different domains are evaluated for testing the feasibility of ASR performance in Punjabi.	2. Quantitative: Experiments were performed on the Kaldi toolkit using the Ubuntu operating system.	Punjabi	There are utterances in the training data from 1. 39 speakers (3. 18 female and 21 male speakers.) The test set has utterances spoken by 1. 6 speakers, having 3. 3 male and 3 female speakers. The 2. age group of children for recording data is 8-14 years.	1. Novel. Initially, only children's data was given to the system, acting as baseline data. Corpus is fed to the system for feature extraction using the MFCC technique, thereafter delta and double delta features are computed and LDA+MLLT are applied on features, then output is fed to DNN-HMM model. Further, the tacotron system is implemented and audios generated from tacotron are merged with children's corpus. Two techniques of feature extraction have been implemented which are MFCC and MFCC with Pitch features. Robustness of the system has been seen when pitch features are used along with MFCC features.

26	Jacob, A.; Mythili, P.	2008	This paper discusses the implementation details of a child friendly, good quality, English text-to-speech (TTS) system that is phoneme-based, concatenative, easy to set up and uses little memory.	2. Quantitative. A comparison of the two methods (concatenation vs LPC), in terms of performance as well as the resources used, has been done using MOS.	English	The recordings were done in male and female voices. 1. A female voice, sampled at 22.05 kHz, was played back at a reduced sampling frequency of 17.5 kHz in order to produce a distinct male voice. **The listeners:** were all non-native speakers of English. There were five teachers (one of them well versed in linguistics), two 8-year-old kids (who are used to synthetic voices), two doctors, and 1 person who had recovered from a voice loss recently.	1. Novel. Speech was produced from the synthesizer by replacing the pitch signal calculated by the function with a vector of constant values. The constant values, that were selected, were 100 Hz and 380 Hz, which correspond to male and child voices, respectively. In the first method, a smooth, direct waveform concatenation of phoneme segments had been done, maximizing the speech quality in terms of naturalness and intelligibility. In the second method, LPC had been used to reduce memory requirements.
27	Jia, N.; Zheng, C.; Sun, W.	2020	With the help of the concept of affective computing, this paper analyses the relationship between emotion and speech from the speech signals of known emotional state, and applies these emotional features to the speech synthesis process, so as to obtain a natural and friendly synthetic voice with rich mood changes and able to simulate human emotions.	2. Quantitative: Experimental design.	Unknown.	The researchers recruited a total of 1. 400 volunteers (2. 50% male and 50% female 3. 5-12 years old, average age 9 years old).	1. Novel. In the experiment, the original speech is used as the baseline, and the following comparison models are designed. Model 1 uses HMM model, model 2 uses conditional GAN model, and model 3 uses current model, CycleGAN. The similarity between the generated speech and the target emotion in the three models is verified.
28	Jreige, C.; Patel, R.; Bunnell, H.T	2009	This project leverages a person's residual phonatory abilities in order to build an adaptive text-to-speech synthesizer that is intelligible, yet conveys the user's vocal identity. Their VocaliD system combines the source characteristics of the disordered speaker with the filter characteristics of an age- matched healthy speaker using voice transformation techniques, in order to produce a personalized voice.	2. Quantitative: Experimental design. Intelligibility and similarity testing.	English	1. Twenty-four monolingual English listeners were recruited to transcribe 220 synthetic sentences created by transforming the vocalizations of 1. four target users with age and gender matched healthy speakers. The four target users consisted of 4. two healthy children (2. 1M, 1F: 3. mean age 10.8 years) and 4. two children with speech impairment due to cerebral palsy (2. 1M, 1F: 3. mean age 11.1 years). Thus, eight transformed voices were generated using combinations of gender matched and unmatched target and control speakers. An additional group of 24 monolingual English listeners were recruited for the similarity test.	1. Novel. VocalID: The speech database is produced by an age and gender matched healthy speaker. These productions are then used to create a blended database that imprints the spectral and prosodic properties of the fluent speech onto source information derived from the target speaker.

29	Karhila, R.; Sanand D. R.; Kurimo, M.; Smit, P.	2012	This paper describes experiments in creating personalised children's voices for HMM-based synthesis by adapting either an adult or child average voice.	2. Quantitative: Experimental design	Finnish	For the listening test, they chose 1. three target speakers randomly: 3. 10- and 11-year old 2. girls and a 9-year old boy out of the test set of ten speakers. They have four different types of average voice models to synthesize the child voice for the above mentioned speakers. This means, they have three sets of synthetic voices for each of the average speaker models and all together 12 synthetic voices for a specific speaker using all the average speaker models. 26 listeners started the listening test.	1. Novel. They present investigations on generating child voices by adapting average voice models. In order to improve the synthetic quality of child voices using an adult average voice model, they propose an approach called stacked transformations. They show that the proposed approach provides very good synthetic voices for children when compared to directly adapting the average voice model. They also investigate the use of vocal tract length normalisation with stacked transformation to assess whether it can provide any improvement when there is very little adaptation data available from the target speaker.
30	Koul, R.; Clapsaddle, K. C.	2006	The purpose of this study was to examine whether individuals with intellectual disabilities demonstrate improvement in the perception of synthetic speech as a result of repeated exposure to it. Koul and Clapsaddle (2006) investigated the effects of practice for the ETI-Eloquence synthesizer.	2. Quantitative: Experimental. Rather than repeating or transcribing the produced synthesized speech, these investigators chose a closed-set picture identification task to derive a dependent measure.	Unknown	1. 28 (18 4. with mild-to-moderate intellectual disabilities; 3. age range 22-55 years and 10 4. typically developing children; age range 4-8 years old)	2. Commercial synthesis system. The synthetic speech used was the ETI-Eloquence (2000) adult male voice Wade.
31	Koul, R.; Hester, K.	2006	To examine the perception of synthetic speech by individuals with severe intellectual disabilities using a closed-response format task.	2. Quantitative. Experimental design. Participants were instructed that they would hear a series of words or sentences preceded by a carrier phrase and that their task was to point to the picture on the computer screen depicting the word.	Unknown	1. 28 (14 with 4. severe intellectual disabilities; 3. age range 13-54 years and 14 4. typically developing children; age range 2;6-3;8 years old)	2. Commercial synthesis system. The synthetic speech used was DECtalk Betty, a simulated adult female voice.

32	Kumar, J.V.; Surendra, A.K.	2011	Using Hidden Markov Model-based system (HTS), a Statistical Parametric Synthesizer is built from child speech. This type of synthesizer has been successfully used previously for such imperfect speech with a limited amount of data. Here, they compare the speaker-dependent and speaker-adaptive models whose outputs are plotted against the natural and the vocoded speech.	2. Quantitative: Experimental design. They evaluate these systems for similarity of the target speaker, naturalness and intelligibility. They also evaluate the vocoded speech against the synthetic speech, considering the difficulties in extracting the fundamental frequency (F0), spectrum estimation and vocoding the child speech.	South Indian English	English Speech of 1. one South-Indian 2. female child of 3. 7-years is recorded in an informal manner. 4. Typically developing.	1. Novel. They use an HMM-based system for child speech synthesis. Speaker-adaptive and Speaker-independent systems are the two types of speech systems that have been designed and evaluated.
33	Liao, H.; Pundak, G.; Siohan, O.; Carroll, M.; Coccaro, N.; Jiang, Q.; Sainath, T.; Senior, A.; Beaufays, F.; Bacchiani, M.	2015	In this paper the authors describe the use of a neural network classifier to identify matched acoustic training data, filtering data for language modeling to reduce the chance of producing offensive results. They also compare long short-term memory (LSTM) recurrent networks to convolutional, LSTM, deep neural networks (CLDNN).	1 Quantitative. Experimental	US English	The generic training sets were obtained by sampling from their Voice Search traffic. The data was then anonymised and hand transcribed. An obvious method to improve children's speech recognition was to collect a matched training set in the same manner. To account for children's vocal tract lengths, they collected wide-band audio so reliable estimation of high frequency filter banks could be made.	1 Novel. The paper compare long short-term memory (LSTM) recurrent networks to convolutional, LSTM, deep neural networks (CLDNN).
34	Matassoni, M.; Falavigna, D.; Giuliani, D.	2016	This paper describes an approach for adapting a DNN trained on adult speech to children voices. The method extends a previous one, based on the Kullback-Leibler divergence between the original (adult) DNN output distribution and the target one, by accounting for the quality of the supervision of the adaptation utterances.	1 Quantitative. Experimental	Italian	The experiments described below were conducted using two different sets of acoustic models (AMs): one trained on an Italian corpus of adult speech, containing broadcast news (IBNC), and the other trained on an Italian corpus of children speech (ChildIt)	1 Novel. The researchers lead DNN adaptation experiments in two different conditions: "age-adaptation" and "self-adaptation". In the first case they use a small amount of adaptation utterances, distinct from the test utterances and transcribed both manually and automatically, to re-estimate parameters of the original DNN. In the second case they perform DNN parameters adaptation using the same test set they were trying to recognize, together with a corresponding automatically generated supervision. Basically, the latter approach extends the previously mentioned speaker adaptation method, based on fMLLR transformations, by adding in the decoding stage a DNN adaptation step.

35	Matassoni, M.; Gretter, R.; Falavigna, D.; Giuliani, D.	2018	This work deals with non-native children's speech and investigates both multi-task and transfer learning approaches to adapt a multi-language Deep Neural Network (DNN) to speakers, specifically children, learning a foreign language.	1 Quantitative. Experimental	Non-native speakers: (a) Italian students, speaking both English and German, and (b) German students speaking English.	In this work the authors exploited speech data collected within the European funded project PF-Star (2002-2004). During the PFStar project, noticeable amount of speech data were collected from English, German, Italian and Swedish children [28]. For the purposes of this work, children's speech pronounced by English, German and Italian students in the three languages were considered.	1 Novel. The researchers use transfer learning to adapt the multi-lingual DNN trained on native data from Italian, German and English children. Basically, only the weights of the output layer of the network are updated, through back propagation, using data from non-native speakers of a given language while the weights of the lower layers are frozen and remain unchanged during adaptation. In addition, they propose to use multi-lingual data even in the adaptation phase, that is to update the weights of the output layer of the original DNN with all available non-native data.
36	Metallinou, A.; Cheng, J.	2014	This work explores the use of DNNs for ELL children's speech recognition. The authors investigate the use of DNNs to improve acoustic modeling of their challenging dataset that contains accented, child speech. They further explore how ASR improvements translate to better modeling of the linguistic content of students' responses through use of latent semantic analysis (LSA), and more accurate subsequent feature extraction for fluency and pronunciation modeling.	1 Quantitative. Experimental	English	The ELL data used in the study was obtained from a large language assessment project administered in schools in a U.S. state. The Arizona English Language Learner Assessment (AZELLA) is an ELL test administered in Arizona to K-12 students.	1 Novel. Their DNN-based speech recognition system, built using rectified linear units (ReLU), greatly outperformed recognition accuracy of Gaussian mixture models (GMM)-HMMs, even when the latter models were trained with eight times more data.

37	Mills, T; Bunnell, H.T; Patel, R	2014	The present study sought to determine whether (a) the source transfer process significantly reduced TTS intelligibility, and (b) identity- bearing source features of the target (speech impaired) talker were preserved in the resultant TTS voice.	2. Quantitative.	English	The surrogate talker (whose fluent speech provided the basis for voice conversion) was an 3. 11-year-old 4. typically developing 2. female. Synthetic voices were generated for four target talkers to assess the voice conversion-synthesis process. Target talkers were two speakers with 4. severe dysarthria (DYS; a 9-year-old male with flaccid dysarthria and a 9-year-old female with mixed spastic-flaccid dysarthria) and two approximately age-matched neurotypical talkers (Non-DYS; an 8-year-old male and an 8-year- old female). To test intelligibility, 21 adult monolingual English-speaking listeners were recruited. A group of 21 adult, monolingual, English-speaking listeners separate from those used in the intelligibility study, were recruited to test the voice conversion element of the VocaliD process.	1. Novel. The ModelTalkerVersion 1 TTS system was modified for the present implementation of the VocaliD process. ModelTalker is a concatenative synthesis system that extends diphone concatenation in two ways. First, the basic storage units are biphones (a pair of adjacent phonetic segments) rather than diphones, and multiple instances are stored for each biphone. Second, splice points within each segment are determined at the time of synthesis when multiple biphone units are considered for each successive pair of phonetic segments, and diphone splice points are determined based on a Viterbi search that minimizes a multivariate cost function based on factors such as F0, spectral disjuncture, and temporal distortion. The ModelTalker version adapted for this project used PSOLA (Pitch Synchronous Overlap and Add) for waveform synthesis, allowing some control of both timing and F0 in the output synthesis.
38	Mousa, A.	2011	Speech morphing was used to modify a source speaker sound to appear like another one's sound.	2. Quantitative.	Unknown.	Speech morphing is one application of speech processing which is usually defined as the modification of speech signal of 1. one speaker (source speaker) to sound as if it has been pronounced by a different speaker (target speaker).	1. Novel. Speech morphing. Different experiments for changing the female to female, male to male, male to female and female to male speech were performed. Also, Child 1 to Child 2, Child 1 to Child 3, Child 2 to Child 3 and Child 3 to Child 2.
39	Murphy, A.; Yanushevskaya, I; Chasaide, A. N.; Gobl, C.	2020	This paper describes the results of a voice transformation task experiment conducted as part of the evaluation of a speech synthesis system (the GlórCáil system). This paper describes a system that allows for the transformation of voices by using our understanding of the speech production process, and in particular, by using an acoustic glottal model-based excitation that more closely models the natural glottal source.	2. Quantitative: Experimental design	Irish English	The baseline stimulus for the voice transformation task was based on a recording of an all-voiced sentence 'We were away a day ago' spoken by 1. one 2. male Irish English speaker. Each of the three target speakers (man, woman, child) was combined with three target affects (sad, angry, no emotion). The baseline sentence was thus to be manipulated by the participants nine times to produce nine (3 speakers x 3 affects) utterances. Listening task: 1. 16 participants, all native speakers of English, took part in the test.	1. Novel. Voice transformation. The GlórCáil system is an application developed for the analysis and resynthesis of speech with a particular focus on control of voice source parameters.

40	Přibilová, A.; Přibil, J.	2006	This paper introduces two new simple non-linear methods of frequency scale mapping for transformation of voice characteristics between male and female or childish.	2. Quantitative: Experimental design. Listening tests. Three types of listening tests have been performed for evaluation of synthetic male, and derived female, childish and young male voices: determination of voice type, naturalness of synthetic speech (MOS), and better sound of transformed voice (pairs of audio samples generated by two non-linear conversion methods were compared).	Czech and Slovak	Three new diphone speech inventories corresponding to female, childish and young male voices are created from the 1. original 2. male speech inventory. 1. Twenty nine listeners within the age from 3. 23 to 75 years (18 Czechs and 11 Slovaks, 2. 17 men and 12 women) took part in the listening tests.	1. Novel. It uses cepstral description of the diphone speech inventory of the male speaker using the source-filter speech model or the harmonic speech model. The text to be synthesized is entered as the input sequence of phonemes. It is converted to the combination of diphones and phone units through the prosody generator. Each of these units has its cepstral description in the speech database. For the real-time TTS synthesis based on the source-filter speech model, the minimum-phase cepstral coefficients are used to model the vocal tract transfer function. For unvoiced speech, the synthesis filter is excited by the noise generator. For voiced speech, a superposition of the impulse generator output and the high-pass filtered noise according to the spectral flatness measure SF is used.
41	Pucher, M.; Toman, M.; Schabus, D.; Valentini-Botinhao, C.; Yamagishi, J.; Zillinger, B.; Schmid, E.	2015	In this paper the researchers evaluate how speaker familiarity influences the engagement times and performance of blind school children when playing audio games made with different synthetic voices.	2. Quantitative: Experimental design	German	They built voices of 1. 18 school children and 7 teachers in the same school and an additional speaker who was not known to the children.	1. Novel. Using an HMM-based speech synthesis system: When developing a synthetic voice for a speaker, they train a separate model for F0, spectrum, and duration for that speaker. These parameters are predicted for each speech unit by taking a large context into account. This leads to a more similar voice than only modifying certain speech parameters like overall duration, F0, frequency range, and loudness.

42	Qian, M; McLoughlin, I; Guo, W; Dai, L	2016	This paper explores children's ASR using baseline DNN-HMM and GMM-HMM systems based on the CMU Kids Corpus. Using this baseline, (i) the researchers evaluate training data augmentation for children by adding noise (which is known to work for adult ASR training but has not yet been evaluated for children), (ii) they introduce a novel training selection approach based on gender and (iii) trial the use of VTLN in opposing directions, since the two directions are currently unexplored for children's ASR.	2 Quantitative. Experimental	US English	CMU Kids Corpus contains about 9 hours of recordings of material read by children. They focus on the larger amount of data from children aged 6 to 9. The researchers also use adult speech from the TIMIT corpus for training. This consists of recordings from 630 male (m) and female (f) speakers of eight American English dialects reading phonetically rich sentences.	1 Novel. DNN-HMM and GMM-HMM systems. Gender selection, VTLN and data augmentation by adding noise.
43	Qian, Y,; Wang, X; Evanini, K; Suendermann-Oeft, D.	2016	In this paper, the researchers explore how to build an automatic speech recognizer for non-native children's speech using a corpus of children's speech of moderate size assisted by a large corpus of adult speech.	2 Quantitative. Experimental.	Non-native English	Two corpora of non-native spontaneous English drawn from the domain of spoken English proficiency assessment are used in this study. The first corpus contains non-native children's speech (children corpus) drawn from a pilot version of the TOEFL Junior Comprehensive assessment (ages 11-15 years old, from over 10 native language backgrounds). The second corpus (adult corpus) is drawn from a large-scale standardized spoken language proficiency test, TOEFL iBT, which measures a non-native speaker's ability to use and understand English at the university level. The adult corpus contains over 800 hours of non-native spontaneous speech covering over 100 native languages across 8,900 speakers.	Novel. o Baseline: DNN-based speech recognizer with i-vectors. The training data used for this system is the AsrTrain partition of the child corpus. o Adp-Adult: Speaker adaption of DNN with i-vectors. The training data used for this system is the AsrTrain partition of the adult corpus. None of the data from the children's corpus is included. o Adp-Comb: Combining children's and adult training data. The AsrTrain partition of the children's corpus and 10,000 responses (~150 hours) randomly selected from the AsrTrain partition of the adult corpus are used as training data for this system. They also tried adding more adult speech into the training set. o Adp-Pretr: DNN pre-training with Adult data. The same data as in the Adp-Comb system is used for acoustic modeling except that the weights of the DNN are initialized by pre-training with the full AsrTrain set from the adult corpus.

44	Saheer, L.; Yamagishi, J.; Garner, P.; Dines, J.	2013	This paper shows that the merits of VTLN can be combined with those of linear transform-based adaptation in a hierarchical Bayesian framework, where VTLN is used as the prior information. The experiments are performed on matched and mismatched train and test conditions.	2. Quantitative: Experimental design. Objective evaluations based on the mel-cepstral distance (MCD) was carried out. The subjective evaluations for naturalness (MOS scores), speaker similarity and intelligibility were performed on three different systems: VTLN, CSMAPLR and CSMAPLR (VTLN prior).	UK English	For the objective evaluations, speech utterances taken from the corpus were used to adapt gender dependent average voice models, which were trained using speech data uttered by about 1. 17 adult 2. male or 19 adult female speakers, respectively. The speaker dependent male model is used as the base model for adapting to the child speech. 1. 14 listeners participated in the subjective evaluations.	1. Novel. (VTLN, CSMAPLR and CSMAPLR (VTLN prior)). A novel technique for propagating the gender and age information captured by the VTLN transform into constrained structural maximum a posteriori linear regression (CSMAPLR) adaptation is presented.
45	Serizel, R.; Giuliani, D	2014	Recent works have shown that augmenting the inputs of a DNN with, e.g. an estimate of the background noise or utterance i-vector, can improve the robustness and speaker independence of the DNN. These researchers propose to augment the MFCC inputs of the DNN with the posterior probabilities of the VTLN-warping factors to improve robustness with respect to inter-speaker acoustic variations.	2 Quantitative. Experimental.	Italian	The APASCI speech corpus: high quality, acoustic-phonetic Italian corpus (194 adult speakers for a total durations of 7h:05m). The ChildIt corpus is an Italian, task-independent, speech corpus that consists of clean read speech from children aged from 7 to 13 years, with a mean age of 10 years.	1 Novel. This paper introduces approaches based on vocal tract length normalisation (VTLN) techniques for hybrid deep neural network (DNN) - hidden Markov model (HMM) automatic speech recognition when targeting children's and adults' speech. VTLN is investigated by training a DNN-HMM system by using first mel frequency cepstral coefficients (MFCCs) normalised with standard VTLN. Then, MFCCs derived acoustic features are combined with the VTLN warping factors to obtain an augmented set of features as input to a DNN. In this later, novel, approach the warping factors are obtained with a separate DNN and the decoding can be operated in a single pass when standard VTLN approach requires two decoding passes.

46	Serizel, R.; Giuliani, D.	2016	This paper introduces deep neural network (DNN)–hidden Markov model (HMM)-based methods to tackle speech recognition in heterogeneous groups of speakers, including children. We target three speaker groups consisting of children, adult males, and adult females.	1 Quantitative. Experimental	Italian	For this study, the authors relied on three Italian speech corpora: the ChildIt corpus consisting of children's speech, the APASCI corpus and the IBN corpus consisting of adults' speech.	1 Novel. Two different kind of approaches were introduced here: approaches based on DNN adaptation and approaches relying on vocal-tract length normalisation (VTLN). First, the recent approach that consists in adapting a general DNN to domain/language specific data is extended to target age/gender groups in the context of DNN–HMM. Then, VTLN is investigated by training a DNN–HMM system by using either mel frequency cepstral coefficients normalised with standard VTLN or mel frequency cepstral coefficients derived acoustic features combined with the posterior probabilities of the VTLN warping factors. In this later, novel, approach the posterior probabilities of the warping factors were obtained with a separate DNN and the decoding could be operated in a single pass when the VTLN approach required two decoding passes.
47	Shahnawazuddin, S.; Dey, A; Sinha, R.	2016	To address the pitch induced distortions, the authors explore pitch-adaptive signal processing for the front-end speech parameterization. Furthermore, the authors have also explored the existing dominant feature-space normalization approaches in the context of the proposed pitch-robust features.	1 Quantitative. Experimental	British English	The ASR system is developed on the WSJCAM0 speech corpus. This database consists of 15.5 hours from 92 adult male/female speakers for training. In the mismatched test scenario, the effectiveness of the explored approaches is evaluated using the children's speech test set (ChTs) of the PF-STAR British English speech database.	1 Novel. For all the experimental evaluations presented in this paper, a speaker independent (SI) ASR system is developed using the Kaldi speech recognition toolkit, employing the DNN based acoustic modeling.
48	Shivakumar, P. G.; Georgiou, P.	2020	This work attempts to address the key challenges using transfer learning from adult's models to children's models in a Deep Neural Network (DNN) framework for children's Automatic Speech Recognition (ASR) task evaluating on multiple children's speech corpora with a large vocabulary. The paper presents a systematic and an extensive analysis of the proposed	2. Quantitative: Experimental design. They perform carefully designed experiments to prove their initial hypothesis that (i) transfer learning from adult ASR is advantageous, and (ii) adapting bottom layers of	US English	In this work they employ 5 different children speech databases and 1 adult speech corpora.	1. Novel. In this study, the researchers conduct an analysis of LVCSR adaptation and transfer learning for children's speech using multiple databases. They compare the advantages of DNN acoustic models over the GMM-HMM systems. Transfer learning from Adults to children- They perform transfer learning from two different base models: (i) Adult's model, (ii) Combined model (Adult + Children) and compare with (iii) Children's ASR to verify their hypothesis for the need for transfer learning from adults to children. Acoustic

			transfer learning technique considering the key factors affecting children's speech recognition from prior literature.	a DNN helps to address acoustic variability.			variability modeling—They conduct experiments on TEDLIUM (Adult's speech) corpus by artificially creating acoustic feature level variability. They apply (i) pitch-shifting, (ii) time-stretching on the raw speech signals and (iii) random VTLN warping on the MFCC features.
49	Shivakumar, P.; Potamianos, A.; Lee, S.; Narayanan, S.	2014	This paper presents a preliminary study towards better acoustic modeling, pronunciation modeling and front-end processing for children's speech.	2 Quantitative. Experimental.	English	Three children speech databases were used in this work: 1)The Children's Interactive Multimedia Project (CHIMP): The data consists of verbal interaction of children ranging between 6 years and 14 years with the computer. 2) The CU Read, Prompted Speech Corpus: Children through grade 1 to 5 (6 years to 11 years) reading sentences and isolated words. 3) CU Story Corpus: The CU Story Corpus consists of read and summarized stories from children ranging from 3rd through 5th grade. 4) Speech data collected from the joint effort of Southwestern Bell Technology Resources and Central Institute for the Deaf (CID): CID consists of five sentences read out by 436 children (5 - 18 years) and 56 adults (25 - 50 years).	1 Novel. Kaldi was configured to model Hidden Markov models (HMM). The researchers evaluate the importance of Cepstral Mean and Variance Normalization (CMVN) and Vocal Tract Length Normalization (VTLN) techniques. Acoustic model adaptation: The researchers experiment with the effectiveness of both speaker independent and speaker dependent techniques. They use MLLT as a standard for speaker independent acoustic adaptation. The speaker adaptive training (SAT) incorporated in their system is based on Constrained Maximum Likelihood Linear Regression (CMLLR).
50	Sinha, R.; Ghai, S.	2009	In this work, the researchers have studied the effect of pitch variations across the speech signals in context of automatic speech recognition.	2. Quantitative. Experimental	English	TIMIT database	1 Novel. In this work, the effect of pitch variations across the signals on MFCC feature has been studied. To begin with, the speech signals from the TIMIT database belonging to low (75-100 Hz) and high (200-250 Hz) pitch ranges were selected. The pitch of the signals was estimated using the ESPS tool available in the Wavesurfer software package. The steady portions of different vowels present in the selected signals were extracted and their corresponding MFCC features were computed.

51	Tong, R; Chen, NF; Ma, B.	2017	These researchers compare two data augmentation approaches for children's acoustic modeling.	2 Quantitative. Experimental.	Mandarin	Native Mandarin Corpus - adult (adult-cn): Native Mandarin Corpus - children (kids-cn): The speakers are 256 gender balanced primary school students of ages 7-12. Non-native Mandarin Corpus (kids-sg): The corpus consists of speech recorded from 255 students aged in 7-12.	1 Novel. In this paper, the researchers compare two data augmentation approaches for non-native children's Mandarin mispronunciation detection. The first approach is a straightforward data combination method, it disregards the child and adult categories and consolidates the two datasets together as one entire set for children's acoustic modeling. The second approach is multi-task learning (MTL) [17] using deep neural networks (DNN).
52	Tong, R; Wang, L; Ma, B.	2017	In this paper, the researchers exploit two deep neural network-based transfer learning techniques for children's speech recognition. The objective is to take advantage of the existing large adult's speech corpora in children's acoustic modelling.	2 Quantitative. Experimental	Mandarin	Native Mandarin Corpus - adult (adult-cn): Native Mandarin Corpus - children (kids-cn): The speakers are 256 gender balanced primary school students of ages 7-12. Non-native Mandarin Corpus (kids-sg): The corpus consists of speech recorded from 255 students aged in 7-12.	1 Novel. The first transfer learning approach is to perform acoustic adaptation on top of the well-trained adult model using children's speech. During the adaptation process, the adult model parameters are optimized toward children's acoustic characteristics with relatively small amount of children's speech. The second approach is multi-task learning (MTL). A shared model is trained from multiple tasks and the model parameters for different tasks are estimated with task dependent objective functions. Different from the acoustic adaptation approach, in the first step, the acoustic characteristics of adults and children are jointly learned through shared hidden layers of a deep neural network; in the second step, two different softmax layers are obtained by optimizing adults and children respectively. They further study the transfer learning on multilingual data. Speech data in a different language is incorporated into the transfer learning process.

53	Tulsiani, H.; Swarup, P.; Rao, P.	2017	It is the goal of this study to consider scalable technology solutions that facilitate oral reading assessment in situations where access to language teachers is limited. The researchers choose the specific context of second language English which is a curriculum subject across schools in rural India where the medium of instruction is primarily the regional language. The authors describe their preliminary efforts to assess the recorded read speech by children in English using an Automatic Speech Recognition (ASR) system.	1 Quantitative. Experimental	English (India)	The authors have collected a total of 1000 recordings spanning 18 stories from 70 students in grades 4-8 of a tribal school in Aine, Mumbai where tablet-based story reading is a scheduled and supervised activity conducted in the school hours as part of the Learn English Through Stories (LETS) project. The CSLU Kids' speech corpus was also used, despite it being American English speech. The TIFR Hindi dataset was used for modelling the non-English phones which are not present in the CSLU Kids' corpus. The GIE corpus, recorded at their lab, was useful for incorporating phonetic realizations of English phones by Indian speakers uttering TIMIT prompts.	1 Novel. The authors focused on building robust acoustic models from mismatched datasets and designing LM to detect mispronunciations effectively for the task of automatic reading assessment.
54	Umesh, S.; Sinha, R.; Sanand, D.	2007	In this paper, the authors address the problem of using vocal tract length normalisation (VTLN) to improve the recognition performance of children's speech. In this paper, they report experiments that suggest that the signal-processing should be done differently for children when compared to adult speech.	1 Quantitative. Experimental	Unknown	Hypothetical reference speaker.	1 Novel. The authors study the effect of filter-bank on the VTLN performance.
55	Vaz, M.; Brandl, H.; Joublin, F.; Goerick, C.	2009	These researchers aimed to endow a robot with a child-like voice which was concretized with the use of a new vocoder-like technique for speech synthesis. They goal was learning them through interaction.	2. Quantitative.	German	Utterances spoken by 1. one 3. 10 year old 2. male child, from the TIDIGITS corpus	1. Novel. They use an acoustic production model derived from the channel VOCODER. It uses a gammatone filter bank in the ERB scale, which offers an optimal trade-off between spectral and temporal resolution, enabling them to synthesize high- (women's and children's) and low- (men's) pitch voices with similar naturalness and intelligibility.

56	Von Berg, S.; Panorska, A.; Uken, D.; Qeadan, F.	2009	The intelligibility, likeability, and speech presentation rates for two types of popular synthesized speech, DECtalk and VeriVox, were investigated across four age groups ranging from 6–85 years old.	2. Quantitative: Experimental design. a) Number of correct words repeated. b). Number of correct sentences repeated. c). Number of correct sentence verifications. d). Likability rating on 5-point Likert-type Scale. e). Rate of speech preference. Three types of statistical analyses were performed for the three main research problems: intelligibility, likeability, and speech presentation rate.	English	Participants were 1. 20 2. females and 19 males divided among 3. four age groups: elementary and secondary school children aged 6–13 years (children); high school and college students aged 14–24 years (young adults); working adults aged 25–65 years (adults); and older adults aged 66–85 years (older adults). These participants represent a broad range of individuals who may be listening partners for those who use a text- to-speech device.	2. Commercial synthesis system. We use two commercially available systems, with three different conditions. Three DECtalk voices were used: Paul (DM), Betty (DF), and Kit (DC). These voices were selected because they represented the standard male, female, and child voices. Three VeriVox (Cepstral) voices were also used: adult male Michael (VM), adult female Sarah (VF), and child Jamie (VC).
57	Watts, O.; Yamagishi, J.; King, S.; Berkling, K.	2009	The purpose of the study was to compare the potential offered by HMM-based adaptation and GMM-based voice conversion techniques for the transformation of existing synthesisers trained on adult speech to the voice characteristics of one child target speaker.	2. Quantitative: Experimental design. Results were evaluated through Blizzard- style listening tests. An XAB test was conducted in which a pairwise comparison was made of the four systems in terms of the similarity of the synthetic speech to the natural speech of the target speaker.	The North American-accented English	This is an HMM-based English voice based on 1. four 2. male and two female voices. A smaller corpus of child speech data was used to adapt the adult model. The North American-accented English speech of 1. one 3. 7-year old tri-lingual (Spanish, English, German) 2. female was used. The synthesiser used as the 'source speaker' in the voice conversion systems (Systems C and D) was an existing unit selection voice which had been built with the data of the speaker SLT from the ARCTIC database. The voice had been constructed using the Multisyn voice building tools.	1. Novel. An existing statistical parametric synthesiser is transformed using model adaptation techniques, informed by linguistic and prosodic knowledge, to the speaker characteristics of a child speaker. This is compared with the application of voice conversion techniques to convert the output of an existing waveform concatenation synthesiser with no explicit linguistic or prosodic knowledge. Both systems adopted the gender-mixed average voice from the HTS entry in the Blizzard Challenge 2007.
58	Wendt, O.; Hsu, N.; Simon, K.; Dienhart, A.; Cain, L.	2019	The goal of this study was to evaluate the efficacy of the iPad-based intervention for increasing functional communication in adolescents and young adults with ASD.	2 Quantitative. Experimental design.	Unknown.	Participants were 1. three individuals 4. with ASD (3. aged 14, 16 and 23).	2. Commercial synthesis system. An iPad®2 with the SPEAKall! application was used as a SGD in this study.