A Systematic Review on Automatic Speech Recognition and Its Advanced Technologies for Children

Leena G Pillai
Research Scholar
Department of Computer Science
University of Kerala

D.MuhammadNoorul Mubarak
Head,
Department of Computer Science,
University of Kerala

ABSTRACT

Automatic Speech Recognition (ASR) is a burgeoning field of research for more than five decades. ASR for adult's speech has conquered many signs of progress, and its application penetrated exceptionally into our daily life. Hence, the ASR for children is striving to turn up from its infant stage. The acoustic and linguistic variability of children makes their ASR a challenging and complex one. In this work, a systematic review conducted on ASR technologies experimented on children in the age group of 3 to 14. The published works in the period 2009-2019 considered for the review.The review started with baseline ASR technologies and concluded with general ASR with different adaptive and robust technologies. The systematic work selection strategies followed in this work are participants, speech corpus, features, methods applied, objective, and outcome of the research.

Keywords: Automatic Speech Recognition, Challenges in Children ASR, Speech Processing, Spectral and Temporal features, Acoustic Modelling, Language Modelling.

1. INTRODUCTION

Human being considered speech as their prominent communication medium. Therefore, the demand for Automatic Speech Recognition (ASR) technology interventions, even in daily life, is rising day by day. Recently, drastic progress achieved in the field of high ASR applications. Most of this research is targeting adult speakers. The researchers are still trying to identify and regulate the developing stage speech recognition challenges of children's utterances and incorporate effective adaption technologies to generalize the model (Potamianos et al., 1997). The developmental changes in children create age-dependent speech signal variability in spectral and temporal features. This variability is the prime hurdle in the development of robust ASR for children (Potamianos et al., 2003). Another challenge is the scarcity of speech corpus. Due to unavailability of Corpus, the researchers forced to limit their subject area according to the existing corpus (Chen et al, 2011).

This work considered the published research papers in the area of ASR for children in the year 2009 – 2010. The speech recognition accuracy of a child utterance is strongly

correlated to some of his features – age, gender, fundamental, frequency, and height. Elenius et al (2014) proved that age and accuracy are strongly correlated. Initially, the baseline ASR conducted on a speech corpus of children applied for evaluating the result with different experiments on training and classification models to enhance the result. Later most of the work highlighted the limitation of children's speech corpus (Cosi Piero et al., 2009). Consequently, studies started on generalizing the ASR by adaption and normalization technologies. This technology enables the inclusion of widely available adult's corpus in training the speech recognition model for children. The adaption performed on pitch variability and normalization conducted on vocal tract variability. The children's speech processing and analysis are emerging areas of research for the development of education and language development application. The core theme of this review is to elucidate the ASR progress, applications and approaches in the areas of children.

2. OBJECTIVE

The purpose of this work is to briefly review the progress achieved in the area of ASR in children's utterances (age: 3-14) in the span of 2009-2019. This paper summarizes the recent state-of-knowledge in the prescribed area. This paper intended to create an understanding for the readers by discussing the findings presented in the included research papers.

3. SYSTEMATIC SELECTION STRATEGY

This work strictly followed some eligibility criteria for paper selection. The works considered those works which deal with speech processing and recognition of children. As the contribution of a research is more prevalent, some of the repeated patterns, clumsy and irrelevant papers have evaded. The paper inclusion criteria are as follow:-

3.1 Participants: - The children in the age group of 3 to 14 have considered for this review (Table 1). Children belong to the age group can be classified into 3-7 (3<=age>7) which may not have much gender variation features, 7-12 (7<=age>12) which can be defined as an intermediate age, and 12 – 14 which shows some of the gender-based features (Bockletet., 2008). Although the 3-7 age groups frequently have stable characteristics, their ambiguous utterances make their recognition difficult. The utterances need to be collected from each age group as well as from boys and girls separately.

3.2 Speech Corpora:- The speech corpus covered in this review are FAUAibo (Schuller et al.,2009), FBK ChildIt (Cosi Piero, 2009; Serizel Romain and Giego Giuliani, 2014; Giuliani Diego and Bagher Baba Ali, 2015), TIDIGITS (Ghai Shweta and Rohit Sinha,

2010), CMU Kids (Qian Mengjie et al.,2016; Kumar Manoj et al.,2017; Fei Wu et al., 2019), CHIMP (Kumar Manoj et al.,2017), OGI (Kumar Manoj et al.,2017), PF-STAR (Matassoni Macro et al., 2018; Sinha Rohit and Shahnawa Zuddin 2018; Dubagunta Pavankumar et al., 2019; yadav et al., 2019), CSLU Kids (Fei Wu et al., 2019). Zourmand et al, (2012) and Rahman et al, (2014) conducted study on Malay language utterances and they created their own regional language corpus. Nisimura R et al, (2011) created own Japanese speech corpus to developed a web based voice interface (Table 1). To create a speech corpus for children is much complex than adults (Kraleva Radoslava, 2016)

3.3 Feature Extraction: - Mel Frequency Cepstral Coefficient (MFCC) is the most commonly used Cepstral feature extraction technique. ASR results in better accuracy with MFCC features. 13-dimensional features are extracted from the MFCC algorithm from one frame. MFCC, along with its regression coefficients (delta and delta delta) enhances the performance of ASR. Another feature extraction applied are Perceptual Minimum Variance Distortion less Response (PMVDR) spectrum estimator (Ghai Shweta and Rohit Sinha, 2010), Fundamental and Formants frequencies (Zourmand et al., 2012), and acoustic and phonetic features such as amplitude, pitch, duration, etc.,(Rahman et al., 2014). In recent time, most of the work focused on adaptive technologies which allow the inclusion of adult's utterances in children ASR. The adaptive and normalized methods covered in this work are Cepstral Mean Subtraction (CMS), Vocal Tract Length Normalization (VTLN), Cepstral Variance Normalization (CVN), Linear Discriminant Analysis (LDA), Maximum Likelihood Linear Transform (MLLT), Variational Mode Decomposition (VMD) and Pitch-Adaptive Cepstral Truncation (PACT) (Table 1).

3.4 Methods and Models: -Figure 1 illustrate the general structure of an ASR. Two models Acoustic Model (AM) and Language Model (LM) is depicted in the general architecture. Some of the ASR consists of one more model Phonetic Model (AM). The model construction as the demand of the problem at hand. Most of the ASR systems have AM and LM. Models are created by a training process in which the speech signal features and their corresponding transcriptions are learned.

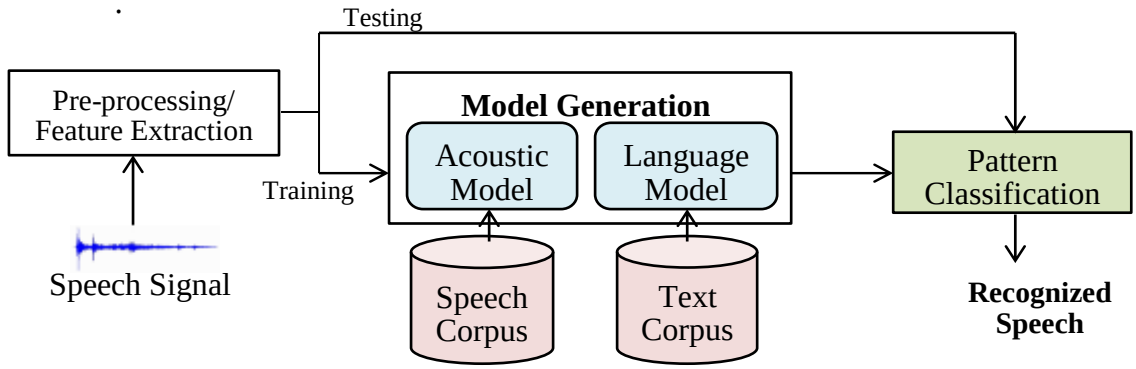

Figure 1: General Architecture of ASR

All the reviewed work used Hidden Markov Model(HMM) to create the Acoustic model. Structured Maximum a Posteriori Linear Regression(SMAPLR), a Bayesian version of transformation-based adaption, applied for fast HMM adaption (Cosi Piero, 2009; Sinha Rohit and Shahnawa Zuddin 2018). Fei Wu et al, (2019) done a camparitive study between baseline Factored Time Delay Neural Network (TDNN) and Time Delay Neural Network (TDNN-F). The work considered done their classification using Guassian Viterbi algorithm (Schuller et al.,2009; Cosi Piero, 2009; Ghai Shweta and Rohit Sinha, 2010; Rahman et al, 2014), Support Vector Machine (SVM) (Nisimura R et al, 2011), Multi-Layer Perceptron (MLP) (Zourmand et al., 2012), Deep Neural Network (DNN) (Serizel Romain and Giego Giuliani, 2014; Giuliani Diego and Bagher Baba Ali, 2015; Qian Mengjie et al.,2016; Kumar Manoj et al.,2017; Matassoni Macro et al., 2018; Sinha Rohit and Shahnawa Zuddin 2018; yadav et al., 2019; Fei Wu et al., 2019), and Convolutional Neural Network (CNN) (Dubagunta Pavankumar et al., 2019) (Table 1).

3.5 Outcome and Discussion

This work tried to track the technical progress record of last ten years (2009-2019) in the area of ASR for children. Detailed result and brief discussions of this review are listed in the Table 1. Before the interventions of DNN in ASR, most of the training and classification are done by HMM and Viterbi decoding respectively. From the table it is clear that, 2014 onwards most of the works are implemented in DNN. HMM is adequate for small vocabulary as well as limited complexity tasks (Schuller et al., 2009; Cosi Piero., 2009; Ghai Shweta and Rohit Sinha, 2010). In large vocabulary and more complex task, some of the extensions, such as GMM (Gaussian Mixture Model) – HMM (Qian Mengjie et al.,2016), DNN – HMM (Serizel Romain and Giego Giuliani, 2014; Giuliani Diego and Bagher Baba Ali, 2015; Sinha Rohit and Shahnawa Zuddin 2018; Yadav et al., 2019). Dubagunta Pavankumar et al, (2019)

designed a CNN architecture that classifies the raw speech samples and proved better even in the adaptive model.

Most of the work followed the same data corpus that may be because of public data scarcity. There are only two regional languages ASR for children were available. By identifying the limited scope of research with available corpus, most of the researches turned towards adaptive technologies. The VTLN combined adaptive technologies are most commonly used speaker adaption technology with better performance. Yadav et al, (2019) experimented in LDA (Noise adaption method), and PACT (pitch adaption method), in adults and children combined model.

ISSN : 0950-0707

Table 1: Review table

Ref:	Participants		Corpora	Features	Method	Outcome
	Age	Utterances				
Schuller et al., 2009	10-13	48401 word utterances from 51 children with 10 emotion such as neutral, joyful, emphatic, irritated, angry, motherese,	German FAU Aibo Emotion Corpus	MFCC coefficients and first and second order regression coefficients.	ASR engine based on continuous Hidden Markov Model. Acoustic Model constructed with 41 phonemes. Three state Phonetic model with five Gaussian mixtures per state.	Emotion recognition from speech. The Accuracy altogether attained is 67.7%. It can be improved with speaker adaption to 76.9%. Best recognition occurred in Emphatic and Angry speech, followed by Neutral, and least accuracy obtained by Motherese speech. String Kernel replaced to vector space modelling and shows better result.
Cosi Piero., 2009	7-13	171 children (85 girls and 86 boys), native sentence utterances from the region on north Italy	FBK ChildIt	Recognizer uses – MFCC Adaption uses- CMS, VTLN, CVN	Recognizer - 3 state HMM with Viterbi alignment frames. Adaption Techniques – SMAPLR	Adults to children adaptive system using SMAPLR with VTLN feature adaption after 5 iteration shows – 25.4% on 40 AUs and 18% on 33 AUs. Italian Children's speech recognition system on children's own corpus using SMAPLR with VTLN and SAT after 5 iteration shows – 18.6% on 40 AUs and 12.2% on 33 AUs
Ghai Shweta, and Rohit Sinha, 2010	6 – 15	11.4 hours of 77 digit sequences records from 326 speakers	TIDIGITS	PMVDR, MFCC and their first and second order temporal derivatives	Digit recognizer developed using HTK Trained with Adults data set (pitch 70-250 Hz). Tested and comparison made with both adults (pitch 80-260 Hz)and children dataset (pitch 100-360 Hz)	Analysed pitch robustness. Default PMVDR shows slightly larger WER (11.57%) than the default MFCC (11.37%). After normalization PMVDR shows the WER of 11.03% and MFCC 9.64%. The LP order changed to 10-25 to optimize PMVDR. PMVDR feature in 15 LP order obtained better result in children test set (9.34%)
Nisimura R et al., 2011	0 – 60, classified to 12 age group, 0-5, 6-10,…, 56-60	2361 short sentence utterances from 1050 web users.	Project own dataset. Large scale Japanese voice collected from 1,152 children through a voice enabled web site and uploaded to server.	12 dimensional MFCC, delta MFCC and delta power	Dataset classified to children and adults(<=16<). HMM with 3-state Gaussians of 128 mixture used to build Acoustic Model. SVM used for classification.	Developed a prototype system for web based voice interface to identify child users based on ASR. Automatic child classifier with HMM + SVM classification based on the 16 year age threshold has shown an average accuracy of 77.8%, which outperform the human hearing ability (66.8%).
Zourmand et al., 2012	7 – 12	Malay sustained vowels -/a/, /e/, /ə/, /i/, /o/ and /u/ for 5 sec each recorded from 360 Malay	Own corpus, recorded with Shure SM58 microphone and by Gold	Fundamental frequency (F0) and three formants frequency (F1,	MLP used for training and classification of 6 vowels with normalized features and Euclidean minimum distance used to classify vowels with	Developed Malay language ASR. The highest accuracy shown by the feature set F0, F1, F2 and F3. The vocal tract normalized using (Nearey Normalization formula) feature set has shown the average accuracy of 86.67% whereas non-

Table 1: Review table ...continued

		children	Wave software with 20 kHz sampling rate and 16-bit resolution	F2, F3).	each formant and fundamental frequency combination and also normalized and non-normalized feature set.	normalized achieved only 69.24% in Euclidean method. The average accuracy obtained from MLP in different age (7-12) and gender (boy and girl) was 86.53%.
Rahman et al., 2014	9 – 12	390 Malay language short sentences which comprises of 1404 words.	Own dataset recorded by lingWAVES software at sampling rate of 48KHz with 16 bit	Acoustic and phonetic features	5- state topology HMM used to build phone level model. The Viterbi algorithm used for classification.	Malay language ASR for children shows the accuracy at sentence level is 71.51% and 76.70% at word level. The Word Error Rate is 23.30% (Deletion – 1.84%, Substitutions – 18.39% and Insertion – 3.07%).
Serizel Romain, and Diego Giuliani, 2014	7-13	171 children (85 girls and 86 boys), 10h:48m utterances.	ChildIt, APASCI (194 Adult speech corpus)	MFCC and its Zero order coefficient and Vocal Tract Length Normalisation (VTLN)	Hybrid DNN (with 4 layers) – HMM 2 models- Age/gender specific DNN-HMM and General DNN-HMM	Research on heterogeneous speech recognition (children, adults – male and female). In limited training dataset normalized MFCC using VTLN for DNN-HMM model improved the performance up to 20% from the base feature set. A general model (ChildIt+APASCI) has shown a Phone Error Rate of 13.12% which is slightly lower than the children PER (14.10%) and much larger than adult PER (female - 10.89%, male – 8.34%)
Giuliani Diego, and Bagher BabaAli, 2015	7 – 13	171 children, each child read 58 or 65 sentences	ChildIt	13 MFCC and its zero order coefficient. These features further normalized using LDA and MLLT.	The hybrid model DNN-HMM are build using LDA+MLLT+fMLLR features and SAT triphone HMM-GMM. 5-gram Laguage Model used	Performance of DNN-HMM compared with SGMM with different SAT. Without SAT HMM-GMM has shown the WER of 16%. All the models showed improved accuracy after SAT. DNN-HMM better reduced WER (7.9%) than HMM-GMM (11.1%) and SGMM (8.5%) with VTLN+fMLLR SAT.
Qian Mengjie et al., 2016	6-11	5180 utterances of English sentences read aloud by children 24 boys and 52 girls.	CMU Kids and TIMIT (Adults dataset)	13 MFCC features, their corresponding Δ and Δ-Δ coefficients, VTLN normalized features.	5-state HMM used to represent each phone, 3-state HMM used to model silence, noise and short-pauses. Trigram Language Model. GMM-HMM and DNN with 4 hidden layers and each layer with 1024 neurons.	Experiment conducted on noisy children's speech dataset, mixed dataset (children + adults and children + adults-female), children + VTLN normalized Adults features and Vice versa. DNN improves WER over GMM-HMM by 12.6%. Mixed data set (adults + children VTLN normalized) achieved relatively better WER in DNN (17.17%) than GMM-HMM (19.06%)
Kumar Manoj et al., 2017	3 - 14	69895 prompted and spontaneous utterances of isolated words, sentences and digits.	Forensic Interviews, CUKids, CHIMP and OGI	40 dimensional MFCC with 100 dimensional i-vectors	DNN with 5 hidden layers and each layer have 1024 nodes. Network used 6.3M tuning parameter. Trigram Language Model used.	Conducted a study on ASR for children in child adult interaction. The Acoustic Model adaption improved the WER up to 11% and Language Model adaption improved the WER up to 15.1%.
Matassoni	9 – 10	Running word	PF-STAR	13 MFCC	HMM – trained with 13 MFCC	Transfer Learning used to adapt the multi-lingual

ISSN : 0950-0707

Table 1: Review table continuation

Macro et al.,2018		utterances from Italian, German and English native and non-native children (total 600 speakers).	Children's Speech Corpus – created as part of EU FP5 PF_STAR project.	coefficients which are normalized further and estimated with fMLLR transformed coefficients are used as feature set	feature. Language Model – bigram Non-native dataset used to train Acoustic Modelling (AM) test done with both native and non-native dataset. DNN with 1632 output layer used for multi-lingual classification.	DNN trained with native utterances from Italian German and English children. Adaption applied on all multilingual utterances separately and evaluated its performance. Multilingual AM always result better performance. Adapted multi-lingual AM with Italian speakers has shown improved WER in English - 25.6% and in German 8.2% utterances.
Sinha Rohit, shahnawazuddin, 2018	9 – 10	Running word utterances from Italian, German and English native and non-native children (total 600 speakers).	PF-STAR British English corpus (children) and WSJCAM0 British English corpus.	39 MFCC coefficients after LDA dimensionality reduction and de-correlated by MLLT.	STRAIGHT approach used for pitch adaption in MFCC. 3-state HMM used for AM. LM– Bi-gram. GMM, SGMM and DNN based model used for pitch adaptive feature evaluation.	The STRAIGHT approach applied pitch adapted MFCC shown improved WER in GMM (-10.1%), SGMM (-5.07) and DNN(-4%) from the default MFCC. SGMM achieved 2.98% improved WER than GMM and DNN achieved 10.04% improved WER than SGMM. Therefore, DNN with STRAIGHT adapted features has shown the best performance.
Dubagunta Pavankumar et al., 2019	4 – 14	158 speaker's utterances of 14.8 hours in British English.	PF-STAR (children corpus) and WSJCAM0 (adults corpus)	13 dimensional MFCC with their corresponding Δ and ΔΔ coefficients.	GMM-HMM trained with mono-phone, tri-phone, LDA+MLLT and LDA+MLLT+fMLLR+SAT. DNN with 429 dimensional MFCC with 11 frames. CNN trained with raw speech samples.	The CNN based system, trained with raw speech, perform consistently better than GMM-HMM and DNN. CNN attained 11.99% WER which is best reported on PF-STAR corpus. Adaption is also worked well in CNN model. Hidden layers in DNN with 1 and 3, and CNN with 3, 4 and 5 are considered in this study. CNN model with 3 layers achieved the best performance.
Yadav et al., 2019	3 – 14	British English 5067 word utterances from 60 children, 1.1 hour.	PF-STAR (children corpus) and WSJCAM0 (Adult corpus)	3-feature set used:- 1)40 dimensional MFCC after LDA and MLLT. 2) VMD-MFCC 3)PACT-MFCC	3-state HMM used for AM and its observation probabilities are generated with DNN and DLSTM.	HMM trained with WSJCAM0+PF-STAR are tested with noisy data and PF-STAR dataset. Recognition performance of VMD-MFCC is 2.9% better WER than baseline MFCC. In noised test dataset DLSTM-HMM with VMD-MFCC shows better performance than MFCC and PACT-MFCC. In pitch adaption also VMD-MFCC shows 1.38% better WER than PACT-MFCC.
Fei Wu et al., 2019	6 – 11 and 3 – 14	6197 utterances, which is phonetically balanced English Simple words, sentences and digits	CMU Kids and CSLU Kids	13 MFCC features, their corresponding Δ and Δ-Δ	Efficiency of TDNN- F in speech recognition is compared with GMM, GMM+VTLN, TDNN (base model), TDNN-F+VTLN	TDNN – F outperforms the base model and it Shows approximately 26% improvement in WER. TDNN-F demonstrates better recognition accuracy in child speech than that of the alternative models considered in this paper.

CONCLUSION

A systematic review conducted on the nature of work published or available in a period 2009 to 2019, in the area of ASR exclusively for children. The feature vector is the fundamental resource of an ASR and 13 dimensional MFCC, recognized as a most commonly used Cepstral feature extraction technique. This paper analyses that apart from baseline ASR, most of the work experiments conducted on advanced adaptive technologies to enhance the speech corpus of children. Adaptive technologies allow the inclusion of adult's speech corpus in Acoustic Model training. The success rate of Adaptive ASR regulates the challenges of children's speech corpus. The adaptive and normalized methods analyses in this work are CMS, VTLN, CVN, LDA, MLLT, VMD and PACT. The VTLN combined adaptive technologies are most commonly used speaker adaption technology with better performance. Noise adaption method, LDA, and pitch adaption method, PACT, experimented with adults and children combined model and result improved performance. The layered architecture of DNN and its capacity to handle complex task makes it suitable for dealing large vocabulary ASR. As DNN outperform the traditional HMM, 2014 onwards most of the work done with DNN.

The availability of work in regional language ASR for children is very limited. The reason may be the unavailability of the public speech corpus of children. As the children are more fluent in their regional language, researches required to turn towards regional language ASR. Several studies proved that technical intervention in speech and language development can make a drastic improvement in children. The effective speech recognition and analysis tools with suitable interface might help children to improve their verbal, non-verbal, and social communication.

REFERANCES

1). Bocklet, Tobias, Andreas Maier, and Elmar Nöth. "Age determination of children in preschool and primary school age with gmm-based supervectors and support vector machines/regression." In *international conference on Text, Speech and Dialogue*, pp. 253-260. Springer, Berlin, Heidelberg, 2008.

2). Chen, Chih-Chang, Ping-Tsung Lu, Meng-Lin Hsia, Jia-You Ke, and Oscal T-C. Chen. "Genderto-Age hierarchical recognition for speech." In *2011 IEEE 54th International Midwest Symposium on Circuits and Systems (MWSCAS)*, pp. 1-4. IEEE, 2011.

3). Cosi, Piero. "On the Development of Matched and Mismatched Italian Children's Speech Recognition Systems." In *Tenth Annual Conference of the International Speech Communication Association*. 2009.

4). Dubagunta, S. Pavankumar, Selen Hande Kabil, and Mathew Magimai Doss. "Improving Children Speech Recognition through Feature Learning from Raw Speech Signal." In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5736-5740. IEEE, 2019.

5). Elenius, Daniel, and Mats Blomberg. "Comparing speech recognition for adults and children." *Proceedings of FONETIK 2004* (2004): 156-159.

6). Ghai, Shweta, and Rohit Sinha. "Analyzing pitch robustness of PMVDR and MFCC features for children's speech recognition." In *2010 International Conference on Signal Processing and Communications (SPCOM)*, pp. 1-5. IEEE, 2010.

7). Giuliani, Diego, and Bagher BabaAli. "Large Vocabulary Children's Speech Recognition with DNN-HMM and SGMM Acoustic Modeling." In *Sixteenth Annual Conference of the International Speech Communication Association*. 2015.

8). Kraleva, Radoslava. "Design and development a children's speech database." *arXiv preprint arXiv: 1605.07735* (2016).

9). Kumar, Manoj, Daniel Bone, Kelly McWilliams, Shanna Williams, Thomas D. Lyon, and Shrikanth S. Narayanan. "Multi-Scale Context Adaptation for Improving Child Automatic Speech Recognition in Child-Adult Spoken Interactions." In *INTERSPEECH*, pp. 2730-2734. 2017.

10). Matassoni, Marco, Roberto Gretter, Daniele Falavigna, and Diego Giuliani. "Non-native children speech recognition through transfer learning." In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6229-6233. IEEE, 2018.

11). Nisimura R., Miyamori S., Kurihara L., Kawahara H., Irino T. (2011) “Development of WebBased Voice Interface to Identify Child Users Based on Automatic Speech Recognition System”.

12). In: Jacko J.A. (eds) Human-Computer Interaction. Users and Applications. HCI 2011. *Lecture Notes in Computer Science*, vol 6764. Springer, Berlin, Heidelberg

13). Potamianos, Alexandros, and Shrikanth Narayanan. "Robust recognition of children's speech." *IEEE Transactions on speech and audio processing* 11, no. 6 (2003): 603-616.

14). Potamianos, Alexandros, Shrikanth Narayanan, and Sungbok Lee. "Automatic speech recognition for children." In *Fifth European Conference on Speech Communication and Technology*. 1997.

15). Qian, Mengjie, Ian McLoughlin, Wu Quo, and Lirong Dai. "Mismatched training data enhancement for automatic recognition of children's speech using DNN-HMM." In *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 1-5. IEEE, 2016.

16). Rahman, Feisal Dani, Noraini Mohamed, Mumtaz Begum Mustafa, and Siti Salwah Salim. "Automatic speech recognition system for Malay speaking children." In *2014 Third ICT International Student Project Conference (ICT-ISPC)*, pp. 79-82. IEEE, 2014.

17). Schuller, Bjorn, Anton Batliner, Stefan Steidl, and Dino Seppi. "Emotion recognition from speech: putting ASR in the loop." In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4585-4588. IEEE, 2009.

18). Serizel, Romain, and Diego Giuliani. "Vocal tract length normalisation approaches to dnn-based children's and adults' speech recognition." In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pp. 135-140. IEEE, 2014.

19). Sinha, Rohit, and Syed Shahnawazuddin. "Assessment of pitch-adaptive front-end signal processing for children's speech recognition." *Computer Speech & Language* 48 (2018): 103-121.

20). Wu, Fei, Povey D. García-PLP, and S. Khudanpur. "Advances in automatic speech recognition for child speech using factored time delay neural network." In *Proceedings of Interspeech*, pp. 15. 2019.

21). Yadav, Ishwar Chandra, S. Shahnawazuddin, and Gayadhar Pradhan. "Addressing noise and pitch sensitivity of speech recognition system through variational mode decomposition based spectral smoothing." *Digital Signal Processing* 86 (2019): 55-64.

22). Zourmand, Alireza, and Ting Hua Nong. "Vowel Classification of Children's Speech Using Fundamental and Formant Frequencies." In *2012 Fourth International Conference on Computational Intelligence, Modelling and Simulation*, pp. 282-287. IEEE, 2012.