Int. j. inf. tecnol. (August 2021) 13(4):1473–1479
https://doi.org/10.1007/s41870-020-00573-y

ORIGINAL RESEARCH

A stacked auto-encoder with scaled conjugate gradient algorithm for Malayalam ASR

Leena G. Pillai[1] · **D. Muhammad Noorul Mubarak**[1]

Received: 23 September 2019 / Accepted: 17 November 2020 / Published online: 31 March 2021

Abstract Automatic speech recognition (ASR) is entitled to automate natural speech perception and the processing mechanism through analysis in the linguistic and acoustic features of the speech signal. ASR for children is highly challenging due to their developing physical aspects and rapidly changing articulation features. Therefore, ASR for children is still at its infant level. In this work, a stacked multilayer auto-encoder (AE) network is designed for ASR of the Malayalam vowel, articulated by children in the age group of five to ten. The proposed network structured with an unsupervised pre-training followed by supervised training. The pre-training coupled with two layers of sparse auto-encoders and scaled conjugate gradient (SCG) algorithm used for back-propagation. The auto-encoders are used to pre-train the network in an unsupervised (self-supervised) manner with 40,500 features that include Mel frequency cepstral coefficients (MFCC) and its derivatives, spectrogram formants and zero crossing rate (ZCR). In the softmax layer, the pre-trained network retrained in a supervised manner with bottleneck features. Fine-tuning has been applied in the trained network to enhance its performance. The unsupervised and supervised layers are stacked together to form a comprehensive network. The designed network has shown an average accuracy of 97% in training and 89.5% accuracy in the test data-set.

✉ D. Muhammad Noorul Mubarak
dmnmubarak@gmail.com

Leena G. Pillai
leenabelieve@gmail.com

[1] Department of Computer Science, University of Kerala, Thiruvananthapuram, India

Keywords Automatic speech recognition (ASR) · Deep neural network · Unsupervised learning · Mel frequency cepstral coefficients (MFCC) · Scaled conjugate gradient (SCG) · Sparse auto-encoder · Spectrogram · Zero crossing rate (ZCR)

Abbreviations

ASR	Automatic speech recognition
AE	Auto-encoder
MFCC	Mel frequency cepstral coefficients
SCG	Scaled conjugate gradient
ZCR	Zero crossing rate
HMM	Hidden Markov model
ANNs	Artificial neural networks
DBN	Deep belief network
RBM	Restricted Boltzmann machine
MOM	Method of moments

1 Introduction

The speech produced can be defined as a result of a concentrated chain process, in which the larynx or the phonation place is the source of the aerodynamic energy of speech, and the articulatory system determines different properties that shape the created sounds [1].

In the case of children, these properties usually exhibit radical variability. These variabilities are categorized as acoustic and articulation variability. Therefore, ASR for children shows significantly poor accuracy than that of adults. By considering these challenges, Mel frequency cepstral coefficients (MFCC), its derivatives, zero crossing rate (ZCR), and spectrogram Formants are identified as interesting features to construct input vector for the auto-encoder network. In this work, the Network is trained and

tested with 10 Malayalam monophthongs vowel utterances of children in the age group of five to ten.

The automatic speech processing technology automates the natural speech chain communication that consists of the speech production mechanism, transmission process, and speech perception process occurred in the ear and brain of the listener. The networks in ASR widely used to study the speech signal through the relevant acoustic information and its statistical representation. In the 1980s, the speech recognition approaches have lifted its methodology from a straightforward template-based paradigm to a more meticulous statistical framework. In the mid-1980s, the hidden Markov model (HMM) become a popular and leading framework for ASR. Since 1988, both theoretical and experimental work has been conducted to analyze the feasibility of artificial neural networks (ANNs) in statistical speech recognition. Hinton et al., proposed an advanced scheme with a fast, greedy, unsupervised, and layer-wise pre-training algorithm in deep belief network (DBN), in which each layer modeled by a restricted Boltzmann machine (RBM) [2]. Later experiments revealed that auto-encoders [3, 4] or conventional neural networks [5] organized in a similar scheme is suitable to model efficient deep neural network (DNN) framework.

The algorithm employed for training and testing has a major role in network performance. A machine learning algorithm can be described as an optimization algorithm to minimize the global error function. One of the line search method called conjugate gradient method was introduced by Hestenes and Stiefel, in the year 1950 as a most prominent iterative method for linear problem solving [6]. In the 1960s, Fletcher and Reeves enhanced this linear method to nonlinear conjugate gradient method. Even though the upgraded algorithm performs much faster than the steepest descent, the calculation complexity per iteration is relatively high since it required a line search to determine the appropriate step size in each iteration. Moller replaced the line search method with the Levenberg–Marquardt approach and introduced a sub-class of the conjugate gradient method called scaled conjugate gradient (SCG) [7, 8]. The SCG reduces the calculation complexity of the conjugate gradient. In this work, the SCG method used to train the network.

By considering the complexity of children's speech recognition, this work proposes a novel approach that finely coupled with customized unsupervised pre-training followed by supervised training. Two different auto-encoders implemented for unsupervised pre-training, which reproduces the input as a result of the output layer. The trial and error approaches are employed to finalize the regularization parameters. As the identification of interesting features of children's speech is much complex than that of adults, the second AE is used to extract bottleneck features. In this work, MFCC features correlated with its derivatives, ZCR, and Spectrogram formants have shown relatively best performance. These bottlenecks (interesting) features identified by the auto-encoders imply for further network training. After pre-training, a softmax training layer designed for supervised training with a sparsely labeled dataset. To accelerate the performance, a supervised fine-tuning applied to the trained network. These training layers are stacked together to form comprehensive network architecture. The designed network for this work has shown an average training accuracy of 97% and a test accuracy of 89.5%.

1.1 Malayalam vowel classification

The vowel sounds are voiced phonemes with the greatest intensity and each Malayalam vowel length usually lies between 40 to 450 ms. Different vowel qualities are produced primarily by altering the position of the tongue (front-to-back and up-to-down) and the lips configuration (neutral, spread or rounded). Malayalam vowels are classified primarily based on the tongue backness and height. Ten Malayalam monophthongs vowels are considered for this study, its classification is listed in Table 1.

2 Literature review

Speech recognition systems for children are much more complicated than that of adults. Russell, Martin, and Shona conducted a study on different parameters that differentiate children speech from adults [9]. The vocal tract length variability and developing articulations make the ASR complicated for children. This study concluded with the fact that adult's speech recognition systems are not adequate to perform on children. Orozco et al., designed an automatic speech recognition system with scaled conjugate gradient (SCG) to classify the infant cry into two classes—normal and pathological cry [10]. The linear predictive coding (LPC) method used to extract features from the

Table 1 Malayalam monophthongs classification

		Front	Central	Back
High	Short	ഇ i		ഉ u
	Long	ഈ i:		ഊ u:
Mid	Short	എ e		ഒ o
	Long	ഏ e:		ഓ o:
Low	Short		അ a	
	Long		ആ a:	

infant cry. The Neural Network with one hidden layer and 15 nodes used for pattern classification. The system has shown an average accuracy of 85%.

Nidhyananthan et al., conducted a study on various feature extraction techniques as well as modeling methods that might be suitable for speech or speaker recognition with developing vocal apparatus (children) [11]. The feature extraction technique considered in this study are Mel frequency cepstral coefficients (MFCC), zero-crossing peak amplitude (ZCPA) and linear predictive coding (LPC) and the modeling methods are Gaussian mixture model (GMM), hidden Markov model (HMM), generalized fuzzy model (GFM) and artificial neural network (ANN). In this experimental study, speech recognition with the MFCC feature vector has acquired the highest accuracy rate of 85%, whereas LPC feature vector has achieved 82% and ZCPA feature vector has shown the least accuracy of 38%.

Sabu, Kamini, and PreetiRao conducted a revised work of ASR in evaluating the reading skills of children in the age group of ten to fourteen with interactive feedback [12]. MFCC features are used to construct the feature vector and DNN for bottleneck feature extraction. The GMM-HMM model is used to train the network. The performance showed an average word error rate of 3.44%.

Most of the researchers applied ASR technology in children for their acoustic speech evaluation. In 1990, the DRA Speech Research Unit recommended for emerging speech recognition technology in speech and language development of children [13]. Vachani et al., have proposed a deep auto-encoder framework to enhance the feature set extracted from Mel frequency cepstral coefficients (MFCC) to improve the performance of ASR in individuals affected by dysarthria [14]. The classification model combined with deep neural network (DNN) and hidden Markov model (HMM) used for automatic speech recognition of dysarthric people. This work achieved an absolute improvement of 16% with auto-encoder that deals with tempo adaption based representation.

Anand et al., have developed a speech recognition application for visually impaired people with MFCC feature and hidden Markov model (HMM) applied to construct the acoustic model [15]. The system achieved an average accuracy of 75%, and after applying speaker adaption technique, the performance improved to 80%.

There are many research work already conducted and is being conducting in the area of regional language. Hence, very limited works are available for children's speech recognition. Till date, automatic speech recognition for children in Malayalam dataset is not available. Therefore, feature extraction methods and classification network architecture for this work is determined based on experiments. The speech recognition works discussed in literature review sessions are used as reference for this work.

3 Methods and implementation

3.1 Data preparation

Data collection and its processing is the primary task of any machine learning approach. The subject of the training dataset consists of 10 Malayalam isolated monophthongs vowels that have been recorded from 150 children, 65 boys, and 85 girls, in the age group of five to ten. Children are very famous for their articulation error and voice opacity. Therefore, the raw dataset went through an audible quality test and preliminary spectrogram study. The raw dataset collected was 1500 (150 × 100), from that only 1350 samples of 10 elements are pruned for network training. The zoomh4n handy portable recorder is used to record the sounds in 16 bit/44.1 kHz sampling rate. The quality of the training data set influence the performance of the classification. The audio editor tool, audacity, is used for speech enhancement. The spectral noise gate technique with 12 dB range applied for noise reduction in an acceptable range.

3.2 Feature extraction methods

A precise feature vector can be embolden pattern recognition accuracy. 13 MFCC features and its corresponding 13 derivatives, one ZCR feature and three spectrogram formants (F1, F2, and F3) are used to create the feature vector, altogether 30 features (1350 samples × 30 features = 40,500 features). Most of the ASR applications attained reasonable performance accuracy with MFCC. MFCC captures the acoustic and perceptual parameters of the speech signal, and it replicates the perception process (Cochlea) of the human being [16–18]. MFCC have very sensitive Filter Banks that filters the speech sounds similar to Cochlea. In this work, 25 filters are applied. In order to cope up with fluctuating characteristics of speech signals, amplified signals enclosed in 30 ms frames with an overlap of 20 ms.

The ZCR consider each window and counts the total number of time the amplitude of the speech signal crosses through zero. The ZCR determines the voiced and unvoiced signal classification [19–21]. All the vowel sounds in Malayalam have voiced sounds, and the ZCR shows low crossing rate than that of unvoiced. Vowels can be well distinguishable with its spectrogram formants [22–25]. According to vocal tract parameters used for vowel constriction, here considered three parameters F1,

F2, and F3. Figure 1, scatter plot, depicts the relationship of each feature variable against other feature variables. The feature selection policy is described in the Table 4.

3.3 Neural network design and implementation

The Auto-Encoder is a pattern learning approach which imposes the unsupervised training model. This model is best suitable for nonlinear as well as a complex range of problem-solving (audio, image, etc.). Each Auto-Encoder consolidated with a pair of encoder and decoder. The encoder is bound to encode the input into hidden layer representation, and the decoder reproduces the input at the output layer [26–28]. The sparse auto-encoder allows an auto-encoder framework to learn interesting patterns even though the hidden layer neurons are greater than the number of inputs ($h^n >= x^n$) by imposing some additional constraints called sparse constraints. The proposed architecture is described in *Fig. 2*.

In this work, the training data consist of unlabeled vowel features {f1, f2, f3,…, fn} as input and the back-propagation algorithm applied to regenerate the input as output (o(i) = f(i)). The training function is based on the optimization of a cost function that measures the error between the input 'f' and the output 'o'. The appropriate regularizers added to the cost function to control the activation of neurons at each layer. The auto-encoder tries to recognize a hypothesis function hw, b(f) ≈ f on input 'f', using the weight 'w' and bias 'b'. The regularizers act as the central controller in the sparse auto-encoder. The L2 weight regularization parameter controls the effect of L2 regularizer in each encoder. The impact of the sparsityregularizer, which applied in the cost function, is controlled by the sparsityregularizer coefficients. The sparsity proportion desires the sparsity of output from the hidden layer. The linear transfer function 'Purelin' is used in the decoder to regenerate input at the output layer (Tables 2, 3).

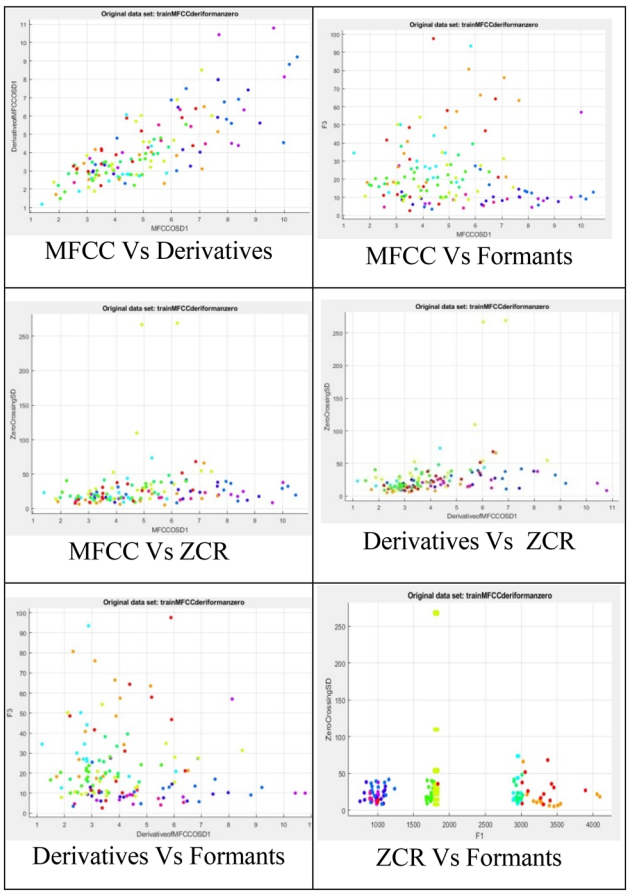

Fig. 1 Relationship between feature variables

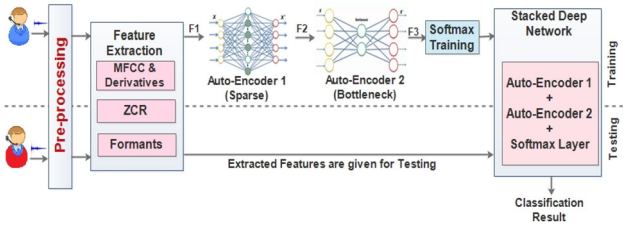

Fig. 2 Relationship between feature variables

Scaled conjugate gradient (SCG): The SCG is a learning algorithm used for neural network. The optimization is the main task performed by the algorithm. The algorithm begins with x_0 and generates a sequence of iterations x_1, x_2, x_3, …, x_k that terminates either by (1) no

Table 2 Auto-encoder 1 parameters

Auto-encoder 1
Parameters
HiddenSize: 100
EncoderTransferFunction: Log Sigmoid
EncoderWeights: [100 × 30 double]
EncoderBiases: [100 × 1 double]
DecoderTransferFunction: Pure Linear
DecoderWeights: [30 × 100 double]
DecoderBiases: [30 × 1 double]
TrainingParameters: [1 × 1 struct]
ScaleData: 1
Training parameter
Training algorithm: SCG
MaxEpochs: 400
L2WeightRegularization: 0.002
SparsityRegularization: 4
SparsityProportion: 0.03
Loss function: MSE

Table 3 Auto-encoder2 parameters

Auto-encoder 2
Parameters
HiddenSize: 28
EncoderTransferFunction: Log Sigmoid
EncoderWeights: [28 × 100 double]
EncoderBiases: [28 × 1 double]
DecoderTransferFunction: Pure linear
DecoderWeights: [100 × 28 double]
DecoderBiases: [100 × 1 double]
TrainingParameters: [1 × 1 struct]
ScaleData: 1
Training parameter
Training algorithm: SCG
MaxEpochs: 100
L2WeightRegularization: 0.004
SparsityRegularization: 4
SparsityProportion: 0.02
Loss function: MSE

more process can be made, (2) the solution has been approximated with sufficient accuracy. The SCG designed by fusing conjugate gradient method and trust region approach (Levenberg–Marquardt algorithm). Therefore, the computation complexity of SCG got reduced significantly.

The network design steps applied for this work are as follows:-

- The first AE designed with 100 neurons. The 30 feature coefficients along with sparsity constraints given to the first auto-encoder. Therefore, output of the first Auto-Encoder is 100 × 1dimension and given as input to the second auto-encoder (Fig. 3).
- The second AE designed with 28 neurons. Therefore, 100 × 1 dimension coefficients compressed to 28 × 1, and this compressed representation considered as bottleneck features for softmax layer training (Fig. 4).
- The softmax layer trained with bottleneck feature extracted from the second auto-encoder in a supervised manner. The sparse matrix used for labelling (10 classes) (Fig. 5).

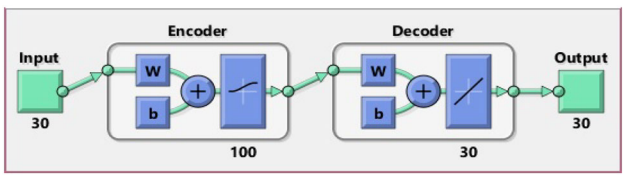

Fig. 3 Auto-encoder 1 architecture

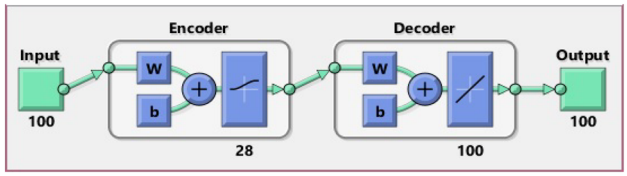

Fig. 4 Auto-encoder 2 architecture

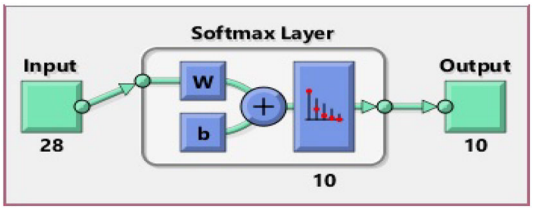

Fig. 5 Softmax-supervised training

- The encoders and the softmax layers are stacked together to form a deep network (Fig. 6).
- Fine-tune the deep network with the labelled dataset.

4 Result and discussion

In this work, an auto-encoder network framework designed for recognizing interesting feature patterns as well as appraising the network performance in classification. This work experimented with different feature extraction methods and their correlations. MFCC features are commonly used parameter for adult's speech recognition system. The study conducted in this work shows that the Malayalam speech recognition system which is exclusively designed for children, achieves relatively low performance with only MFCC features. The feature extraction techniques that contributed to the highest performance—MFCC and its derivatives, ZCR, and spectrogram formants—are considered as suitable features for this work (Table 4).

The architecture consists of two layers of unsupervised or self-supervised learning model, auto-encoder, employed for pre-training and feature identification. The sparsely labeled dataset used for further supervised training and fine-tuning. Fine-tuning is an optional training layer. In this work, the fine-tuning improved the training result from

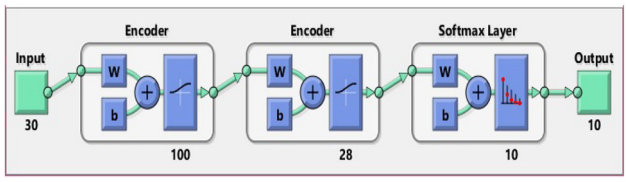

Fig. 6 Stacked auto-encoder

Table4 Feature selection experiment result

Features	Training	Testing
MFCC	68	60
MFCC + derivatives	76	67
MFCC + derivative + formants	90	81
MFCC + derivative + formants + ZCR	97	89.5
MFCC + derivative + formants + ZCR + MOM	88	79

87.7 to 97% and also shows a drastic enhancement in testing as well (Table 5).

The designed network framework showed an average training accuracy of 97% (Fig. 7) and testing has shown an average accuracy of 89.5% on the test dataset (Fig. 8).

The trained network classifies the test data to ten classes such as class 1—അ (a), class 2—ആ (a:), class 3—ഇ (i), class 4—ഈ (i:), class 5—ഉ (u), class 6—ഊ (u:), class 7—എ (e), class 8—ഏ (e:), class 9—ഒ (o), class 10—ഓ (o:). 100% classification accuracy achieved by class 1, 2 and 5. Among the other classes, the worst false positive (FP) rate (19.0%) has shown by class 9 and 10. However, class 6 has shown the least accuracy with the highest false negative (FN) rate (25.0%). The false events (false positive and false negative) can be classify into interclass and intraclass misclassification. According to the articulation required to articulate vowels, the short vowels and its corresponding long vowels [e.g., അ (a) and ആ (a:)] can consider as in one class (Table 1). A most identifiable parameter that distinguishes the long and short vowel in a class is its duration of articulation. The long vowels [e.g., ആ (a:)] duration is higher than the short vowel [e.g., അ (a)]. Most of the speakers are not able to understand the duration difference required for uttering long and short vowel. Therefore, intra-class misclassification in isolated vowels is quite common even in adults. However, intra-class isolated vowels classification errors shall be negligible as their articulation will be corrected when combined with words.

5 Conclusion

Children got highly influenced by technologies. Several studies proved that Assistive Technologies improves speech and language development in children. Automatic

Table5 Importance of fine-tuning

	Training	Testing
Without fine-tuning	87.7	74
With fine-tuning	97	89.5

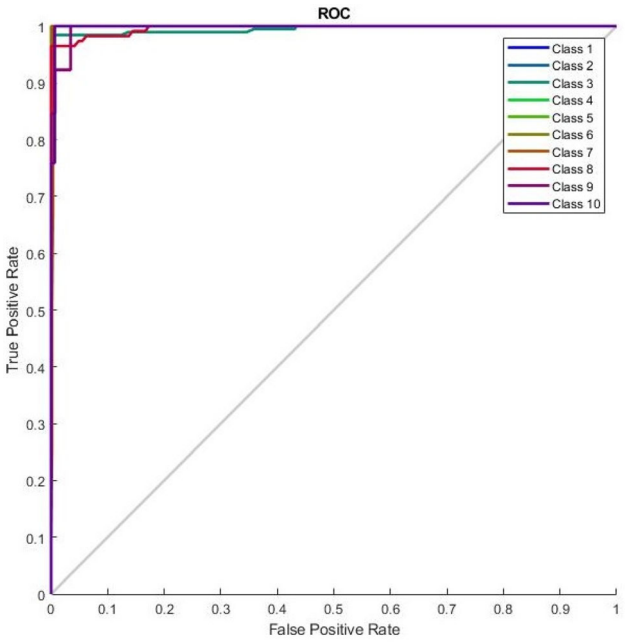

Fig. 7 Training performance

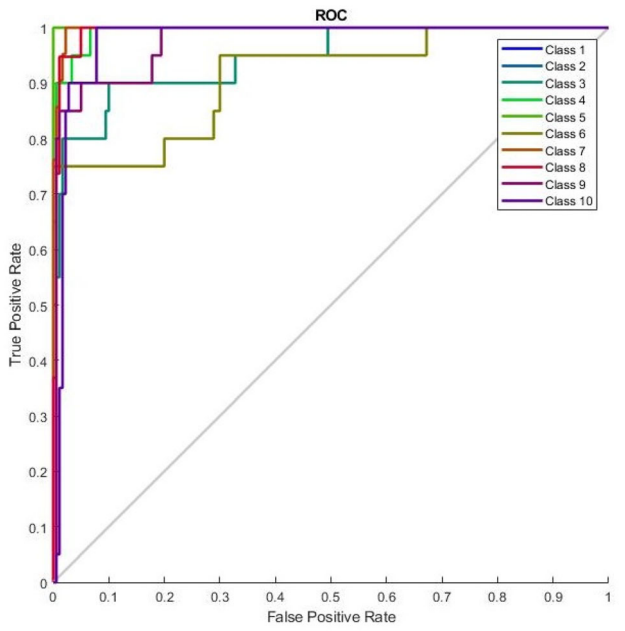

Fig. 8 Performance on test data

speech recognition (ASR) enhances the features of Assistive Technologies. In this work, a neural network framework, which is pre-trained in an unsupervised manner (auto-encoder) and fine-tuned with a sparsely labeled dataset, used to recognize the Malayalam vowels articulated by children belongs to the age group of five to ten. Two auto-encoders used for pre-training with 100 and 28 hidden layer neurons, respectively. The bottleneck features extracted from the second auto-encoder with 28 neurons. The performance of the classification has been accelerated by supervised fine-tuning. In this work, the softmax layer trained with scaled conjugate gradient (SCG) method that combines the conjugate gradient method and the Trust Region method. Therefore, SCG regulates the calculation complexity at each iteration. The ASR for children is much more challenging than that of adults. This work conducted an initial ASR work in Malayalam vowels for children and has shown an average accuracy of 89.5% in the test dataset.

References

1. Ionescu CM (2013) The human respiratory system. The human respiratory system. Springer, London, pp 13–22
2. Hinton GE, Osindero S, Teh YW (2006) A fast learning algorithm for deep belief nets. Neural Comput 18(7):1527–1554
3. Ranzato MA, Huang FJ, Boureau YL, Le Cun Y (2007) Unsupervised learning of invariant feature hierarchies with applications to object recognition. In: IEEE conference on computer vision and pattern recognition, CVPR'07, 2007. IEEE, pp 1–8
4. Pillai LG, Sherly E (2017) A deep learning based evaluation of articulation disorder and learning assistive system for autistic children. Int J Nat Language Comput (IJNLC) 6(5)
5. Deng L, Hinton G, Kingsbury B (2013) New types of deep neural network learning for speech recognition and related applications: an overview. In: 2013 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 8599–8603. IEEE, 2013
6. Hager WW, Zhang H (2006) A survey of nonlinear conjugate gradient methods. Pac J Optim 2(1):35–58
7. Møller MF (1993) A scaled conjugate gradient algorithm for fast supervised learning. Neural Netw 6(4):525–533
8. Khadse CB, Chaudhari MA, Borghate VB (2016) Electromagnetic compatibility estimator using scaled conjugate gradient backpropagation based artificial neural network. IEEE Trans Ind Inform 13(3):1036–1045
9. Russell M, D'Arcy S (2007) Challenges for computer recognition of children's speech. In: Workshop on speech and language technology in education, 2007
10. Orozco J, Reyes García CA (2003) Detecting pathologies from infant cry applying scaled conjugate gradient neural networks. In: European symposium on artificial neural networks, Bruges (Belgium), pp 349–354, 2003
11. Nidhyananthan SS, Shantha Selvakumari R, Shenbagalakshmi V (2014) Contemporary speech/speaker recognition with speech from impaired vocal apparatus. In: 2014 international conference on communication and network technologies (ICCNT), pp 198–202. IEEE, 2014
12. Sabu K, Rao P (2018) Automatic assessment of children's oral reading using speech recognition and prosody modeling. CSI Trans ICT 6(2):221–225
13. Russell M, Brown C, Skilling A, Series R, Wallace J, Bonham B, Barker P (1996) Applications of automatic speech recognition to speech and language development in young children. In: Spoken language, 1996. ICSLP 96. Proceedings, fourth international conference on, vol 1, pp 176–179. IEEE, 1996
14. Vachhani B, Bhat C, Das B, Kopparapu SK (2017) Deep auto encoder based speech features for improved dysarthric speech recognition. Proc Interspeech 2017:1854–1858
15. Anand AV, Shobana Devi P, Stephen J, Bhadran VK (2012) Malayalam speech recognition system and its application for visually impaired people. In: India conference (INDICON), 2012 annual IEEE, pp 619–624. IEEE, 2012
16. Ittichaichareon C, Suksri S, Yingthawornsuk T (2012) Speech recognition using MFCC. In: International conference on computer graphics, simulation and modeling (ICGSM'2012), July, pp 28–29, 2012
17. Kumar AP, Kumar N, Kumar CS, Yadav AK, Sharma A (2016) Speech recognition using arithmetic coding and MFCC for Telugu language. In: 2016 3rd international conference on computing for sustainable global development (INDIACom), pp 265–268. IEEE, 2016
18. Lad NR, Nirmal JH, Naikare KD (2019) Total variability factor analysis for dysphonia detection. Int J Inf Technol 11(1):67–74
19. Kulkarni N (2018) Use of complexity based features in diagnosis of mild Alzheimer disease using EEG signals. Int J Inf Technol 10(1):59–64
20. Shete DS, Patil SB, Patil S (2014) Zero crossing rate and energy of the speech signal of Devanagari script. IOSR JVSP 4(1):1–5
21. Panda SP, Nayak AK (2016) Automatic speech segmentation in syllable centric speech recognition system. Int J Speech Technol 19(1):9–18
22. Bansal S, Agrawal SS, Kumar A (2019) Acoustic analysis and perception of emotions in hindi speech using words and sentences. Int J Inf Technol 11(4):807–812
23. Huber JE, Stathopoulos ET, Curione GM, Ash TA, Johnson K (1999) Formants of children, women, and men: the effects of vocal intensity variation. J Acoust Soc Am 106(3):1532–1542
24. Sainath TN, Mohamed A-R, Kingsbury B, Ramabhadran B (2013) Deep convolutional neural networks for LVCSR. In 2013IEEE international conference on acoustics, speech and signal processing, pp. 8614–8618. IEEE, 2013
25. Ahmad W, Shahnawazuddin S, Kathania HK, Pradhan G, Samaddar AB (2017) Improving children's speech recognition through explicit pitch scaling based on iterative spectrogram inversion. In: INTERSPEECH, pp 2391–2395, 2017
26. Gehring J, Miao Y, Metze F, Waibel A (2013) Extracting deep bottleneck features using stacked auto-encoders. In: 2013 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 3377–3381. IEEE, 2013
27. Hsu W-N, Glass J (2018) Extracting domain invariant features by unsupervised learning for robust automatic speech recognition. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 5614–5618. IEEE, 2018
28. Dendani B, Bahi H, Sari T (2020) Speech enhancement based on deep auto encoder for remote Arabic speech recognition. In: International conference on image and signal processing, pp 221–229. Springer, Cham, 2020