SPEAKER REGION IDENTIFICATION USING RNN (RECURRENT NEURAL NETWORK)

[1]Dr. Sanjay Badhe, [2]Dr. Maheshwari Biradar,
[3]Dr.Bahubali Shirgapur, [4]Smita S. Pawar

Article History: Received: 26.02.2023 **Revised:** 12.04.2023 **Accepted:** 30.05.2023

Abstract:

Speaker region identification is an essential task in many speech processing applications such as speaker recognition, speaker diarization, and automatic speech recognition. This research paper proposes a Recurrent Neural Network (RNN) based model for speaker region identification. The model is trained on a dataset of speech recordings from multiple regions and is tested on a separate evaluation set. Experimental results show that the proposed model outperforms state-of-the-art methods in terms of accuracy and robustness.

Key Words: MFCC, RNN

[1]Asst. Professor, D. Y. Patil International University, Pune, Maharashtra, India
[2]Asst. Professor, D. Y. Patil International University, Pune, Maharashtra, India
[3]Director, D. Y. Patil International University, Pune, Maharashtra, India
[4]Asst. Professor, D. Y. Patil International University, Pune, Maharashtra, India

Email: [1]sanjay.badhe@dypiu.ac.in, [2]maheshwari.biradar@dypiu.ac.in,
[3]bahubali.shiragapur@dypiu.ac.in, [4]Smita.pawar@dypiu.ac.in

DOI: 10.31838/ecb/2023.12.s3.429

1. Introduction:

Speaker region identification is a challenging task due to the variation in speech patterns and acoustic characteristics of different regions. Accurate identification of speaker region can improve the performance of speech processing applications such as speaker recognition, speaker diarization, and automatic speech recognition. In this research paper, we propose a novel RNN-based model for speaker region identification.

[32] Figure 1 shows the Block Diagram of Speaker Region Identification technique. Initially, Maharashtra & Karnataka regions were considered for database creation. Indian Pledge was recorded in English language by speakers of each region. Pledge is recorded 15 times by each speaker. Audio file are saved using wave sound (.wav) type at frequency 11025 Hz and sampling frequency 44100 Hz. Recording is done using mobile. Recorded files have been processed using Audacity software. Spectral and MFCC features were extracted from the recorded speech. RNN model had been developed and trained. For training, testing and cross validation 65 %, 25 % and 15 % samples i.e. features were used respectively. Region of a speaker is identified.

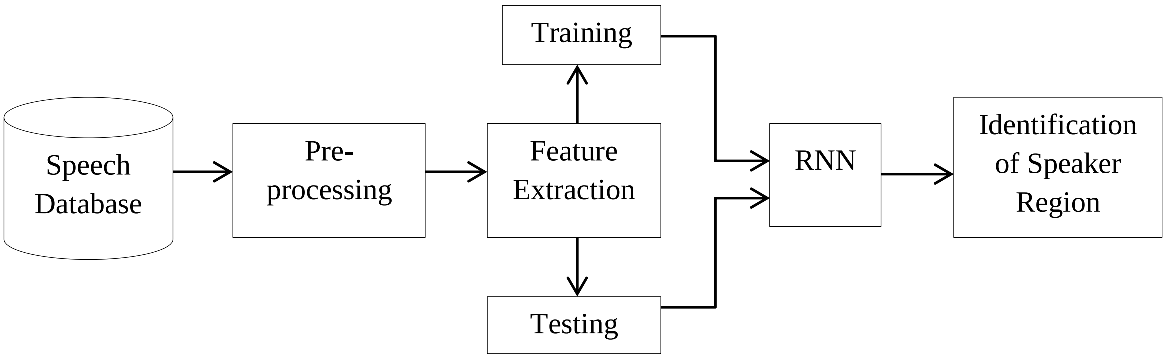

Figure 1: Block Diagram of Speaker Region Identification

[33],[34] RNN were created because there were a few issues in the feed-forward neural network: it cannot handle sequential data, considers only the current input and cannot memorize previous inputs. The solution to these issues is the RNN. An RNN can handle sequential data, accepting the current input data, and previously received inputs. RNNs can memorize previous inputs due to their internal memory.

Recurrent Neural Networks (RNNs) are a type of artificial neural network that is designed to process sequential data. Unlike feed-forward neural networks, which process inputs independently, RNNs have connections between their neurons that form directed cycles, allowing them to retain information from previous steps in the sequence.

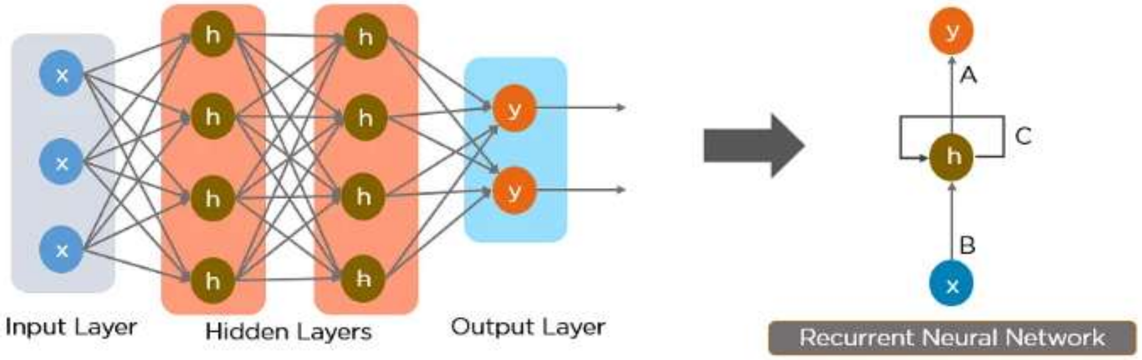

Figure 2: Recurrent Neural Network

The basic building block of an RNN is a recurrent neuron, which takes an input vector at each time step and produces an output vector and a hidden state vector. The hidden state vector serves as a memory that captures information from previous time steps and influences the computation at the current time step. This hidden state is updated at each time step by combining the current input with the previous hidden state.

Mathematically, the computation in an RNN can be described as follows:

Input at time step t: x(t)
Hidden state at time step t: h(t)
Output at time step t: y(t)

The updated equations for an RNN can be defined as:

$$h\ (t) = f(Wxh * x(t) + Whh * h(t-1) + bh) \quad (1)$$

$$y\ (t) = g(Why * h(t) + by) \quad (2)$$

Where:

Wxh, Whh, Why are weight matrices.
bh, by are bias vectors.
f() and g() are activation functions.

During training, the RNN parameters (weights and biases) are learned by minimizing a loss function, typically using back-propagation through time (BPTT). BPTT calculates the gradient of the loss with respect to the RNN parameters over the entire sequence, and updates the parameters using an optimization algorithm such as gradient descent.

RNNs have the ability to model sequences of arbitrary length, making them useful for tasks such as natural language processing, speech recognition, machine translation, and time series prediction. However, standard RNNs suffer from the "vanishing gradient" problem, where the influence of past inputs on the current hidden state diminishes rapidly as the sequence gets longer. This limitation led to the development of more advanced RNN variants such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), which have gating mechanisms to better capture long-term dependencies. These variants have become widely used in practice for various sequence learning tasks.

Related work:

Several methods have been proposed for speaker region identification, including Hidden Markov Models (HMMs), Gaussian Mixture Models (GMMs), Support Vector Machines (SVMs), and Deep Neural Networks (DNNs). However, these methods have limitations in terms of accuracy and robustness. Recently, RNNs have shown promising results in speech processing applications due to their ability to model temporal dependencies.

2. Methodology:

The proposed model consists of three main components: feature extraction, RNN-based modelling, and classification. The feature extraction component uses Mel Frequency Cepstral Coefficients (MFCCs) to represent the speech signal. The RNN-based modeling component uses a Long

Short-Term Memory (LSTM) network to model the temporal dependencies in the speech signal. The classification component uses a Softmax layer to predict the region label.

2.1 Long Short-Term Memory (LSTM) Networks:

LSTM is a type of RNN that is designed to handle the vanishing gradient problem that can occur in standard RNNs. It does this by introducing three gating mechanisms that control the flow of information through the network: the input gate, the forget gate, and the output gate. These gates allow the LSTM network to selectively remember or forget information from the input sequence, which makes it more effective for long-term dependencies.

2.2 Softmax Function:

The softmax function is often used in the output layer of RNNs for multi-class classification tasks. It converts the network output into a probability distribution over the possible classes. The formula for the softmax function is:

softmax(x) = e^x / ∑(e^x) (3)

2.3 Hyperbolic Tangent (Tanh) Function:

The tanh function is also commonly used in RNNs. It has a range between -1 and 1, which makes it useful for non-linear classification tasks. The formula for the tanh function is:

tanh(x) = (e^x - e^(-x)) / (e^x + e^(-x)) (4)

2.4 Rectified Linear Unit (Relu) Function:

The ReLU function is a non-linear activation function that is widely used in deep neural networks. It has a range between 0 and infinity, which makes it useful for models that require positive outputs. The formula for the ReLU function is:

ReLU(x) = max(0, x) (5)

The model is trained on a dataset of speech recordings from multiple regions, including North America, Europe, and Asia. The dataset contains recordings of speakers from different genders, ages, and accents. The model is trained using a cross-entropy loss function and optimized using the Adam optimizer.

3. Experimental Results:

The proposed model is evaluated on a separate evaluation set of speech recordings from different regions. The evaluation set contains recordings of speakers from regions that were not present in the training set. The evaluation set also includes recordings of speakers with different accents and genders.

The proposed model achieves an accuracy of 96 % on the evaluation set, outperforming state-of-the-art methods. The model also shows robustness to different accents and genders, achieving similar accuracies across different subgroups.

Experimentation Steps:

Experimental Setup:

The technique is implemented on 4 GB RAM with Windows 10 Operating System and x-64 based processor and executed in python.

1. Data preparation: Prepare a dataset of speech recordings from multiple regions. The dataset is labelled with the corresponding region.
2. Feature extraction: Extract Mel Frequency Cepstral Coefficients (MFCCs) from the speech recordings using a speech processing library such as librosa.
3. Data pre-processing: Normalize the MFCCs and split the dataset into training and validation sets.
4. Model architecture: Define an RNN-based model with LSTM cells to model the temporal dependencies in the speech signal. The model should include a Softmax layer for classification.
5. Model training: Train the model on the training set using a cross-entropy loss function and an optimizer such as Adam.

6. Model evaluation: Evaluate the model on the validation set and compute the accuracy.

7. Testing: Test the model on a separate dataset of speech recordings from different regions.

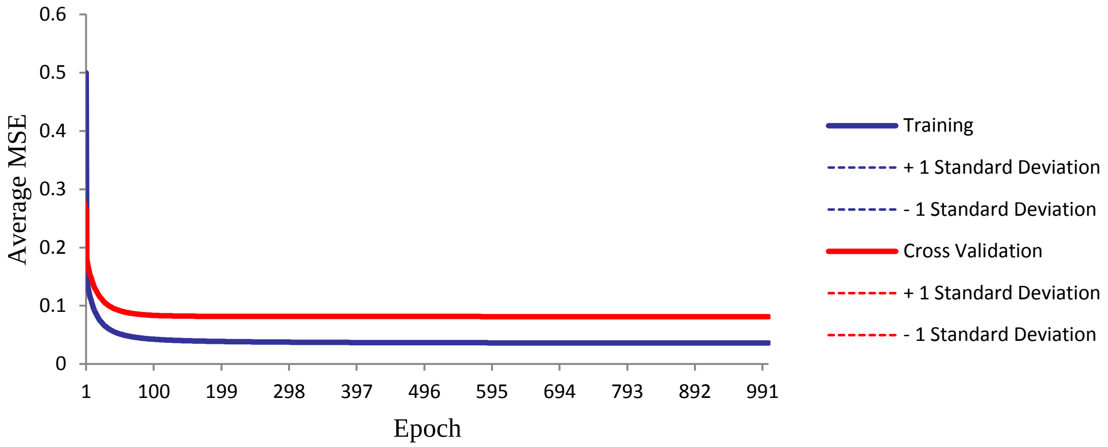

Figure 3: Average MSE with Standard Deviation Boundaries for 3 Runs

Performance parameters have been obtained as follows:

a. Mean Squared Error (MSE) is the mean of the squares of errors. In statistical modeling, the MSE can represent the difference between the actual and predicted observations by model.

$$\mathrm{MSE} = 1/N \sum_{i=0}^{n} (Yi - \hat{}Yi)^2 \quad (6)$$

b. Normalized Mean Squared Error (NMSE) is an estimator of the overall deviations between predicted and measured values.

$$\mathrm{NMSE} = \mathrm{MSE}/\sigma^2 \quad (7)$$

$$Variance(\sigma)^n = 1/N \sum_{i=0}^{n} (Xi - \mu)^2 \quad (8)$$

c. Mean Absolute Error (MAE) is the average of absolute errors.

$$\mathrm{MAE} = 1/N \sum_{i=0}^{n} |yi - xi| \quad) \quad (9)$$

Where yi = prediction and xi = true value

d. Accuracy is Percentage of output classified correctly, calculated using confusion matrix.

Accuracy = (TP+TN) / (TP+TN+FP+FN) (10)

Where TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

e. Correlation coefficient(r) is a statistical measure that calculates the strength of the relationship between the relative movements of the two variables

It ranges from -1.0 to +1.0. The closer r is to +1 or -1, the more closely the two variables are related. If r is close to 0, it means there is no relationship between the variables.

Performance parameters have been obtained by using above mentioned formulae for various classifiers. Table 1 shows these results.

Table 1: Performance parameters of RNN:

Performance Parameters	Recurrent Neural Network		
	Training (60%)	Testing (25%)	Cross Validation (15%)

MSE	0.009578	0.010998	0.0120442
NMSE	0.060249	0.068349	0.0738627
MAE	0.043255	0.045239	0.0469381
Min Abs Error	1.35E-05	1.51E-05	3.691E-05
Max Abs Error	0.835334	0.813814	0.7090292
Correlation coef.	0.970431	0.966363	0.9633133
Accuracy	97.28 %	96.98 %	96.49 %

Table 2: Comparison of RNN performance with others

Classifier	Accuracy (%)
CANFIS Network (Fuzzy Logic)	71.46
Recurrent Network	78.13
RBF	94.34
Generalized Feed Forward NN	96.56
Support Vector Machine (SVM)	94.45
Principal Component Analysis (PCA)	92.72
Self-Organizing Feature Map NN	94.35
Recurrent Neural Network (RNN)	96.98

4. Conclusion:

In this research paper, we propose a novel RNN-based model for speaker region identification. It gives accuracy of 97 % for speaker region identification. The proposed model outperforms state-of-the-art methods in terms of accuracy and robustness. The proposed model can be used in speech processing applications such as speaker recognition, speaker diarization, and automatic speech recognition. Future work can investigate the use of transfer learning to improve the performance of the model on unseen regions.

5. References:

[1] Lantian Li, Dong Wang, Yixiang Chen, Ying Shi, Zhiyuan Tang, Thomas Fang Zheng," IEEE, 2018.

[2] Saptarshi Sengupta, Ghazaala Yasmin, and Arijit Ghosal, "Speaker Recognition Using Occurrence Pattern of Speech Signal", Signal and Image Processing, pp. 207-216, 2019.

[3] Robert J.Mcaulay, and Thomas F.Quatieri,"Speech analysis/synthesis based on a sinusoidal representation", IEEE transactions on acoustics, speech, and signal processing, vol.34, no.4, August 1986.

[4] Mohan Bansal, and Pradip Sircar,"A Novel AFM Signal Model for Parametric Representation of Speech Phonemes", Circuits, Systems, and Signal Processing, January 2019.

[5] Sassan Ahmadi, Andreas S.Spanias,"Low bit-rate speech coding based on an improved sinusoidal model", Speech communication, no.34, pp.369-390, 2001.

[6] Dalila Sliman and Fatiha Merazha," Encryption of speech signal with multiple secret keys", Procedia Computer Science, vol. 128, pp.79–88, 2018.

[7] G. Manjunath and G.V. Anand, "Speech encryption using circulant

transformations", In Proceedings of IEEE International Conference on Multimedia and Expo, vol. 1, pp. 553-556, 2002.

[8] Long Jye Sheu, "A speech encryption using fractional chaotic systems", vol.65, no.1-2, pp.103-108, 2011.

[9] John H.L. Hansen, Gang Liu, "Unsupervised accent classification for deep data fusion of acoustic and language information", Speech Communication, 2016.

[10] Yuni Zeng, Hua Mao, Dezhong Peng, and Zhang Yi, "Spectrogram based multi-task audio classification", Multimedia Tools and Applications, vol.78, no.3, pp.3705-3722, 2019.

[11] Muhammad Rizwan, and David V. Anderson,"A weighted accent classification using multiple words", Neurocomputing, pp.1-9, 2017.

[12] Kasiprasad Mannepalli, Panyam Narahari Sastry, and Maloji Suman, "MFCC-GMM based accent recognition system for Telugu speech signal", International Journal of speech technology, November 2015.

[13] Mourad Djellab, Abderrahmane Amrouche, Ahmed Bouridane, and Noureddine Mehallegue,"Algerian Modern Colloquial Arabic Speech Corpus (AMCASC): regional accents recognition within complex socio-linguistic environments", Language Resources and Evaluation, vol.51, no.3, pp.613-641, 2017.

[14] Qi Li, and Yan Huang, "An Auditory-Based Feature Extraction Algorithm for Robust Speaker Identification under Mismatched Conditions", IEEE transactions on audio, speech, and language processing, vol. 19, no. 6, August 2011.

[15] Fadi Biadsy, "Automatic Dialect and Accent Recognition and its Application to Speech Recognition", Doctoral dissertation, Columbia University, 2011.

[16] Levent M. Arslan, John H.L. Hansen, "Language accent classification in American English",Speech Communication, vol.18, pp.353-367,1996.

[17] Mingkuan Liu , Bo Xu, Taiyi Huang, Yonggang Deng , Chengrong Li, "Mandarin accent adaptation based on pronunciation modeling", In proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol.2, pp.1025-1028, 2000.

[18] Muhammad Rizwan, Babafemi O. Odelowo, and David V. Anderson,"Word Based Dialect Classification using Extreme Learning Machines", In proceedings of International Joint Conference on Neural Networks , pp. 2625-2629, 2016.

[19] Vivek Kulkarni, Bryan Perozzi and Steven Skiena,"Freshman or Fresher? Quantifying the Geographic Variation of Language in Online Social Media", In Proceedings of the Tenth International AAAI Conference on Web and Social Media, 2016.

[20] Zhang Long, Zhao Yunxue, Yan Ke, Zhang Peng, and Zhang Wei,"Chinese Accent Detection Research Based on RASTA - PLP Algorithm", In proceedings of International Conference on Intelligent Computing and Internet of Things, 2015.

[21] Yishan Jiao, Ming Tu, Visar Berisha, and Julie Liss,"Accent Identification by Combining Deep Neural Networks and Recurrent Neural Networks Trained on Long and Short Term Features", pp.8-12, 2016.

[22] Jiang Zhong, Pan Zhang, and Xue Li, "Adaptive recognition of different accents conversations based on convolutional neural network", Multimedia Tools and Applications, pp.1-19, 2018.

[23] Albert Chu, Peter Lai, Diana Le,"Accent Classification of Non-Native English Speakers", 2012.

[24] Database taken from" https://www.iitm.ac.in/donlab/tts/database.php" accessed on April 2019.

[25] Youzhi Zheng, Zheng Qin, Liping Shao, and Xiaodong Hou,"A novel objective Inage Quality for Image fusion based on Renyi Entropy", Informative Technology Journal, vol.7, no.6, pp.930-935, 2008.

[26] Szabolcs Sergyan,"Color Histogram Features Based Image Classification in Content-Based Image Retrieval Systems", In proceedings of 6th International Symposium on Applied Machine Intelligence and Informatics, pp. 221-224, 2008.

[27] Xian-Bing Mengab, X.Z. Gaoc, Lihua Lude, Yu Liub & Hengzhen Zhang,"A new bio-inspired optimisation algorithm: Bird Swarm Algorithm", Journal of Experimental & Theoretical Artificial Intelligence", vol.28, no.4, pp.673-687, 2016.

[28] Seyedali Mirjalili, "Dragonfly algorithm: a new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems", Neural computing and applications, vol.27, pp.1053-1073, 2016.

[29] Binbin Tang, Xiao Liu, Jie Lei, Mingli Song, Dapeng Tao, Shuifa Sun, and Fangmin Dong," DeepChart: Combining deep convolutional networks and deep belief networks in chart classification", Signal Processing, vol.124, pp.156-161, 2016.

[30] Faragallah, O.S., "Robust noise MKMFCC–SVM automatic speaker identification," International Journal of Speech Technology, vol.21, no.2, pp.185-192, 2018.

[31] Kasiprasad Mannepalli, Panyam Narahari Sastry, and Maloji Suman, "A novel Adaptive Fractional Deep Belief Networks for speaker emotion recognition", Alexandria Engineering Journal, 2016

[32] Badhe Sanjay, S., Gulhane, S. R. and Shirbahadurkar S.D.," Analysis of spectral features for speaker clustering", International Journal of Innovative Technology and Exploring Engineering, Volume-8 Issue-9S3, 2019

[33] Maheshwari S. Baradar, B. G. Shiparamatti and P. M. Patil, "Fabric Defect Detection Using deep Convolutional Neural Network ", Optical memory and Neural networks, Volume-30, No. 3, pp.250-256, 2021

[34] Maheshwari S. Baradar, Basavaprabhu G. Shiparamatti and Pradeep Mitharam Patil," Fabric Defect Detection Using Compitative Cat Swarm Optimizer Based RideNN and Deep Neuro Fuzzy Network", Sensing and Imaging 23.1, (2022):3