Journal of Scientific & Industrial Research
Vol. 82, August 2023, pp. 851-860
DOI: 10.56042/jsir.v82i08.3122

Estimating Social Background Profiling of Indian Speakers by Acoustic Speech Features

Mohammad Ali Humayun*, Hayati Yassin & Pg Emeroylariffion Abas

Faculty of Integrated Technologies, Universiti Brunei Darussalam, Jalan Tungku Link, Brunei Darussalam

Received 07 December 2021; revised 20 June 2023: accepted 23 June 2023

Social background profiling of speakers refers to estimating the geographical origin of speakers by their speech features. Methods for accent profiling that use linguistic features, require phoneme alignment and transcription of the speech samples. This paper proposes a purely acoustic accent profiling model, composed of multiple convolutional networks with global average-pooling layers, to classify the temporal sequence of acoustic features. The bottleneck representations of the convolutional networks, trained with the original signals and their low-pass filtered copies, are fed to a Support Vector Machine classifier for final prediction. The model has been analysed for a speech dataset of Indian speakers from social backgrounds spread across India. It has been shown that up to 85% accuracy is achievable for classifying the geographic origin of speakers corresponding to regional Indian languages; 17% higher than the benchmark deep learning model using the same features. Results have also indicated that classification of accents is easier using the second language of the speakers, as compared to their native language.

Keywords: Accent identification, Low pass filtering, Ensemble learning, Native language identification, Speaker profiling

Introduction

Human speech carries substantial information in addition to linguistic content. Automatic estimation of paralinguistic information from speech includes speaker identification and speaker profiling. Speaker identification or verification refers to the identification of a target speaker as matching with one of the speakers from the already available speech samples.[1] On the other hand, speaker profiling represents the estimation of the physical or social characteristics of an unknown speaker. Physical profiling predicts speaker characteristics such as age, height, gender, and weight[2], whilst social profiling predicts race, ethnicity, or homeland of an unknown speaker from the test speech samples. Estimation of social background can be made based on either accent or dialect differences in speech.[3] Accent classification from speech has multiple applications. It can be used to adapt and improve Automatic Speech Recognition (ASR) models, for personalized human-machine interactions and is especially useful in forensics.[4,5] For forensic speaker profiling, the spoken accent is identified to estimate the social background fan unknown speaker from the speech sample available as crime-related evidence, for instance from threat, ransom, or harassment calls. Forensic accent recognition is distinctive as accents to be classified can be from adjoining locations and similar to an extent; that the accent information may not be useful for other applications of accent identification. Furthermore, the available speech sample that is used as evidence may be of very low quality, such that forensic accent recognition may be very challenging.

Both linguistic and acoustic features have been used for the classification of accents. For the linguistic techniques, classification is based on linguistic features, phoneme articulation attributes, or distance among phoneme acoustics. Linguistic features consider the syntactic structure of transcribed words, articulatory attributes represent place and manner of human phoneme articulation, and phonetic distance refers to distance matrix among acoustic features for the set of phoneme utterances.[6,7] Linguistic features may not be applicable in certain circumstances, such as for very short duration speech samples with insufficient phonemes, or low resource languages without text transcription. On the other hand, acoustic accent recognition is usually based only on acoustic features of the speech, which describe the physical characteristics of the sound produced. Prosody and short-term spectral characteristics are usually considered acoustic features for speech processing. Prosodic features

*Author for Correspondence
E-mail: mohammadalihumayun@gmail.com

include information related to pitch, speed, loudness, and pauses in speech, whereas short-term spectral characteristics are inspired by human hearing physiology and are the most widely used features for automatic speech processing.

As discussed further in the next section, promising results in accent classification from similar accents mostly use models utilizing linguistic features, with pronounced deterioration in results for purely acoustic techniques.[5] Purely acoustic models have reported respectable results only for the identification of native language from English speech by speakers from different countries with significant accent and pronunciation differences.[8] Vast geographical coverage of native languages makes their classification inapplicable for applications such as forensic profiling. The few works that have focused on purely acoustic classification for proximate speakers have used the same speakers for training and testing sets, which are split based on the spoken sentences.[9]

This paper aims to analyse the identification of native states from the accent of unseen speakers of different states within India, using the multilingual Indian corpus for speaker profiling.[10] The paper proposes and evaluates the fusion of classifiers using multiple Low Pass Filtered (LPF) versions of the same signal. A convolutional network with global-average pooling, and variants of Recurrent Neural Network (RNN) have been used for performance comparisons, for accent classification tasks using time series of acoustic features. The model has been analysed for speech in both, native and second, languages by the same speakers to record the effect of spoken language on native state classification. Finally, a secondary dataset comprising of different speaker origins has also been used to support the evaluation results for proposed methods.

Literature Review

Most of the research involving the extraction of paralinguistic information from speech, has been traditionally focused on emotion recognition, speaker diarization, and speaker identification.[1–12] The proposed models predominantly convert short-term features of speech extracted from small stationary time windows, into a long-term feature vector, before applying a classifier for discrimination. The fixed-size long-term vector is calculated by statistical modelling over short-term features for complete utterance, and contains information of a longer duration of speech irrespective of the linguistic content. Lately, end-to-end neural network based models have been shown to outperform traditional statistical approach, in capturing long term features for speaker identification tasks.[13,14] Additionally, the estimation of physical and social traits of an unknown speaker has also been established as an attainable task. For physical profiling, Goel *et al*.[15] have developed an application to estimate the speaker's gender, accent, and age using real-time speech. Kalluri *et al*.[2] have demonstrated results for the prediction of speakers' height, weight, shoulder size, and waist size; by utilizing short duration of speech using harmonic and formant features along with Mel-spectrogram. The impact of time window length used for extracting spectral features on the estimation of height, weight, and age, has also been analysed by Rita *et al*.[16] Research for the analysis of accent, dialect, or geographical origin within social profiling has varying contexts. The socio-linguistic community has generally focused on studying the variation in phonetic articulation traits for people from different geographical backgrounds.[17,18] Within the machine learning research, automatic estimation of social background stretches across spoken dialect identification[6], native language (L1) identification from second language (L2) speech[8], and identification of accent within a dialect, for forensics as well as for adaptation of speech recognition models.[5,4,19]

Numerous studies have investigated issues, particularly relevant to accent identification in the forensic context. Brown[3] has tested existing accent recognition models with accents from neighbouring geographic locations, to investigate relative difficulty in forensic applications.[3] In another work[20], the same author has analysed the impact of spoken words and phonemes on classification model accuracy. Brown[5] has also tested spontaneous and degraded speech for similar accents to simulate and analyse realistic forensic accent profiling using existing models. Considering the social aspects of the forensic application for accent profiling, Hughes and Wormald[21] have highlighted humanitarian concerns associated with dependence on technology for investigation and prosecution. To test geographically proximate accents for forensic application, Brown[5] has used the 'Accent and Identity on Scottish English Border (AISEB)' dataset, which has been collected by social science researchers. The dataset consists of speakers from four different towns along the Scottish

English Border, with speech samples categorised as one of the four possible accents. Support Vector Machine (SVM) has been used to classify the accents based on phoneme-dependent York-Accent-Distance (Y-ACCDIST) matrices, whereby the Y-ACCDIST matrix represents the distance among the set of phonemes with the phonemes represented by the average of MFCC coefficients extracted from the midpoint of all utterances by the speaker. The experiments have reported 86.7% accuracy for clean speech and 64.4% after degrading the speech samples to low quality. The same dataset has also been used fortext-independent classification, by utilizing a Gaussian Mixture Model-Universal Background Model (GMM-UBM). Although the result has been shown to be higher than the chance level accuracy of 25% for four target classes, it is still far behind the accuracies obtained using phoneme dependent models.[3] In fact, most models that have been proposed for accent recognition in the literature and have reported good results and they have used linguistic features. Najafian *et al*.[4] have proposed a model that uses phono tactic features along with i-vector, to classify different variants of British English; to adapt and improve ASR accuracy. The 'Accents of British Isles (ABI)' dataset has been used, containing 14 different accents within British English, and the model has reported 84.87% accuracy.[4] Ge[22] has also proposed a linguistic features dependent accent identification model to improve the ASR results. Seven different accents from the Foreign Accented English (FAE) corpus, have been used, with the corpus consisting of English speech by non-native English speakers with significant accent differences. Distance among features for vowel set have been used for classification using GMM-UBM, to achieve a 54% accuracy.[22] On the other hand, Weninger *et al*.[19] have used accent recognition to improve ASR for Mandarin with a purely acoustic model. The model classifies speakers from 15 different regions in China with significant accent variations using i-vectors and Bidirectional Long Short Term Memory (Bi-LSTM). Accuracy of 34.1% has been reported for the accent classification model, higher than chance level accuracy of 6.7%, but still relatively lower than models utilizing linguistic features.

Purely acoustic models have reported encouraging results from the perspective of the speakers' native language identification from speech in their L2 English. Jiao *et al*.[23] have proposed a fusion of Recurrent Neural Network (RNN) and Deep Neural Network (DNN) on the Native Language Speech Corpus (NLSC) in the INTERSPEECH 16 Native Language Sub-Challenge, consisting of TOEFL recordings for English speech by speakers from 11 different countries. The model has reported a 51.92% accuracy. Using the same NLSC corpus, Rajpal *et al*.[8] have proposed a model using MFCC in conjunction with log of phrase-level F0, to identify the native language of speakers, with the model reporting a 40.2% accuracy.[8] Classification of accents for speakers with English as their second language is relatively easier due to the impact of respective native languages on the accents of the speakers, and the typical mistakes they make in English pronunciation. Hence, results for acoustic classification of native language are better than acoustic models for forensic and ASR adaptation applications.

Some purely acoustic accent classification models have reported higher accuracies for similar accents of South Asian languages, however, their experimental setups do not exclude speakers of the test samples from the training set, which makes the setup inapplicable for profiling unseen speakers. Soorajkumar *et al*.[24] have classified 5 different dialects of the South-Indian, Kannada language using neural networks over MFCC, and have achieved 83% sentence-level accuracy. The experiments have used 5 fold cross-validation and have avoided reserving speakers for testing to avoid overlapping with the training set. A neural network model has also been used [9] to identify native province from Urdu speech by speakers from 4 different provinces of Pakistan. The speakers had a regional language as their native and Urdu as their second language. However, the tests have used different speech samples by the same set of speakers for testing and training.

This paper tests accent profiling of unseen speakers using their English and Hindi speech from the multilingual Indian dataset developed by Kalluri *et al*.[10] by fusing multiple Convolutional Neural Networks (CNNs). The proposed model belongs to a category of deep-learning classification models that use purely acoustic features. Most of the acoustic models in the literature have relied on variants of RNN for modelling time series of short-term features.[19,23] Moreover, most fusion techniques have merely merged the output layers by simple averaging or weighted averaging by using empirically adjusted

weights.[22,23] On the other hand, this paper has proposed and analysed the fusion of multiple convolutional networks and found it to give a better performance than standalone RNN or CNN based architectures. The proposed model uses a meta-SVM classifier to fuse the individual classifiers by concatenating the bottleneck layers from the trained end-to-end CNNs and using them as inputs to the SVM classifier. Merging the bottleneck layers has been found to be more effective than the simple fusion of output layers. Performance of the proposed classification method has been compared with the method proposed by Jiao *et al*.[23], which fuses Long Short Term Memory (LSTM) with DNN.

Methodology

This section describes the proposed classification model with the architecture details, as well as the dataset used for evaluation.

The Proposed Classification Model

The proposed classification model is illustrated in Fig. 1. The proposed model extracts acoustic features from three instances of each audio speech sample, to be fed into three parallel 1D convolutional networks having global-average pooling as the final layers. The three instances of the speech sample comprise two LPF versions along with the raw, unfiltered signal. The convolutional networks compress the variable-length acoustic features to a fixed-sized embedding. These bottleneck embeddings from the three convolutional networks are then merged and used to classify the accents using a Support Vector Machine (SVM) classifier.

Acoustic features are the time-series of logarithmic Mel-scaled filter-bank energies, with 39 filter energies extracted from a window size of 25 milliseconds and a hop length of 15 milliseconds. The three speech instances are fed into three parallel classifiers; the first instance is kept intact, whilst the other two instances are passed through fourth-order low-pass Butterworth filters with cut-off frequencies of 1 kHz and 4 kHz, respectively. The signals are then transformed into logarithmic Mel-scaled filter-bank energies, and normalized by the mean and standard deviation of the dataset. Eventually, 39 filters are used for the original as well as the low-pass filtered versions of the speech, increasing the resolution of filters for the lower frequency components in the filtered versions. The higher resolution to focus on lower frequency components is motivated by the fact that most distinctive features in human speech reside in lower frequencies.[25] The three sets of features (one set of features from each instance) are then fed into the parallel convolutional networks.

The Mel filter-bank features from the three instances of the signal are transformed to fix-sized

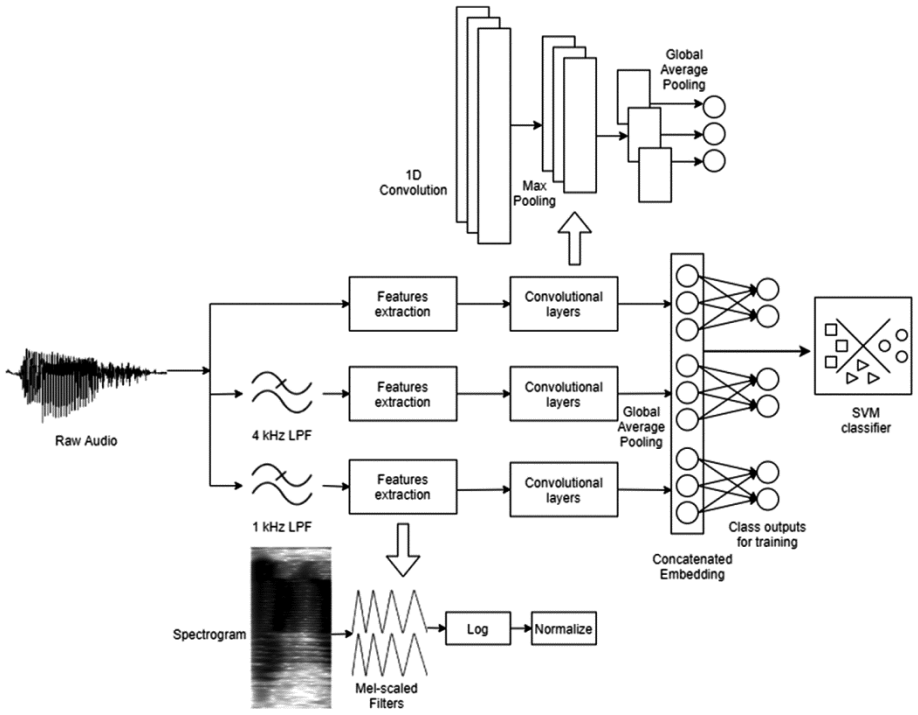

Fig. 1 — The proposed classification model

embeddings by parallel convolutional networks. Each of the convolutional networks consists of three 1D convolutional layers, with each layer having a filter size of 16, kernel size of 8, and a stride length of 1. Each of the convolutional layers is followed by a 4-sample max-pooling. Finally, global-average pooling averages the time axis for each of the filters from the third convolutional layer into a single value and hence, compressing the signal into a 16-dimensional embedding corresponding to the 16 filters. Each of the parallel embedding layers is connected to a dense output with softmax activations to train the end-to-end CNNs for the task of native state classification. The convolutional networks are optimized by minimizing categorical cross-entropy loss using the 'Adamax' algorithm, with the training stopping early when training loss does not decrease for 1 iteration.

After the convolutional networks are trained on the training dataset, they are used to transform copies of the signals onto the 16-dimensional embeddings in their bottleneck layers. The three 16 dimensional embeddings are then concatenated and used as input for the meta SVM classifier with the SVM trained using embeddings by the convolutional networks for the training dataset. The proposed model is then evaluated using the predictions by the trained SVM classifier for the test dataset embeddings as input. The SVM classifier uses a linear kernel.

Dataset

The proposed classification model has been analysed using multiple classification tasks by selecting different groups of speakers as target classes with varying geographical and social spread from the Indian multilingual [10] dataset.

The Indian multilingual dataset contains speech, in English, by Indian speakers having five different native languages: Hindi, Telugu, Malayalam, Tamil, and Kannada. Besides English speech, the dataset also has speech samples by the same speakers in their native languages. From the Indian multilingual dataset, English speech by all the speakers spread across India has been used to evaluate the proposed model. This data subset is subsequently referred to as the 'Indian-English' dataset. The speakers with Hindi as native language represent the largest group in the dataset, hence for analysis across the Hindi region, native Hindi speakers have been further sub-divided into classification groups. For classification within the Hindi region, English speech only by Hindi speakers has been selected as the 'Hindi-English' dataset. Hindi speech, by the same Hindi region speakers, has also been used for analysis and is referred to, as the 'Hindi-Hindi' dataset.

First classification task uses the 'Indian-English' dataset and presents the primary analysis for the proposed model. The speakers have been grouped into 5 target classes according to their native languages (Hindi, Telugu, Malayalam, Tamil, and Kannada) for the 1st classification task. A bulk of the states that have Hindi as their native language has been labelled as a single class named Hindi states. Andhra Pradesh (AP) and Telangana have been grouped as a single class for Telugu whilst Tamil Nadu (TN) and Puducherry have been grouped as the class for Tamil. The classes Kerala and Karnataka correspond to Malayalam and Kannada native languages, respectively. The three speakers in the dataset labelled as Hindi, whilst having Karnataka as native state were dropped from analysis.

Rest of the classifications tasks, i.e.tasks 2–5, represent analysis for the Hindi region only. For the 2nd classification task, the Hindi states have been sub-divided into the Eastern and Western Hindi states; whilst for the 3rd classification task, they have been sub-divided into the Northern and Central Hindi states (2 target classes), using the 'Hindi-English' dataset. Similarly, for the 'Hindi-Hindi' dataset, the Hindi states have been sub-divided into the Eastern and Western Hindi states for the 4th classification task; and into the Northern and Central Hindi states for the 5th classification task (2 target classes). The different classification tasks undertaken in this paper have been summarized in Table 1, with Table 2 giving the classification details of native states for the different classification tasks. The 1st, 2nd/4th, and 3rd/5th classification tasks, have been illustrated in Figs. 2(a), 2(b), and 2(c) respectively, in maps form.

Results

The proposed classification model has been evaluated with the Indian-English, Hindi-English, and Hindi-Hindi datasets, which originated from the

Table 1 — Summary of classification tasks

Task	Speakers	Language	Classes	Class names
1	All	English	5	Hindi, AP, Kerala, TN, Karnataka
2	Hindi	English	2	East, West
3	Hindi	English	2	North, Centre
4	Hindi	Hindi	2	East, West
5	Hindi	Hindi	2	North, Centre

Table 2 — Classes for the different classification tasks

Class	Native states	Classification Task
Hindi states	Rajasthan, Odisha, Madhya Pradesh (MP), Haryana, Chhattisgarh, Uttar Pradesh (UP), Delhi, Maharashtra, West Bengal(WB), Jharkhand, Bihar, Uttarakhand, Panjab, Himachal Pradesh (HP), Jammu &Kashmir (J&K), Gujarat, Meghalaya	1
AP	AP, Telangana	1
Kerala	Kerala	1
TN	TN, Puducherry	1
Karnataka	Karnataka	1
Western Hindi states	Gujarat, Maharashtra, Rajasthan, Panjab, HP, Haryana, J&K, Delhi	2&4
Eastern Hindi states	MP, UP, Odisha, Jharkhand, Meghalaya, WB, Bihar, Chhattisgarh, Uttarakhand	2&4
Northern Hindi states	Rajasthan, Haryana, UP, Delhi, Uttarakhand, Panjab, HP, J&K	3&5

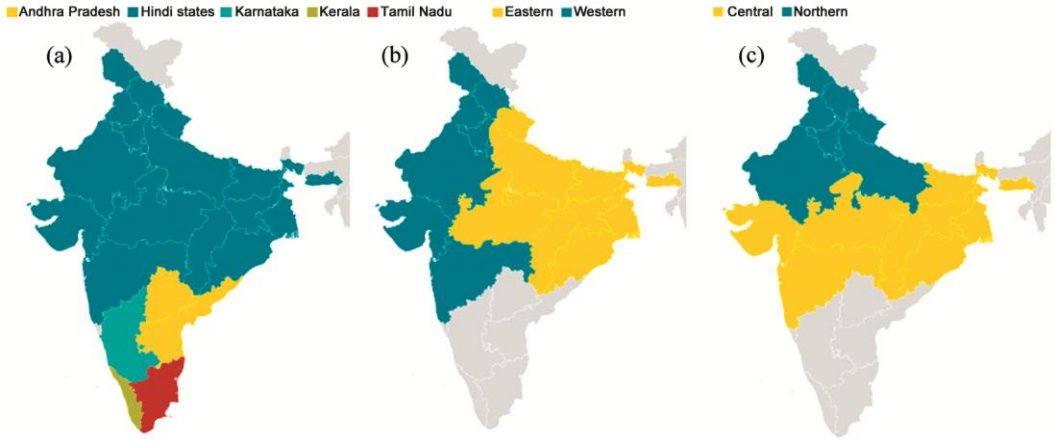

Fig. 2 — Classification tasks illustration: (a) Task 1, (b) Tasks 3&5, (c) Tasks 4&6

Indian multilingual dataset. Multiple Classification Tasks (CTs) have been used for the evaluation of the proposed classification model.

The Indian multilingual dataset repository contains a list of training and test speakers which has been used to obtain the training and test partitions for all the experiments concerning that dataset and results in a total of 206 training and 135 test speakers. A single sentence by each speaker has been used for training and testing.

Evaluation of the Proposed Classification Method

Performance using the proposed classification model has been compared with a single CNN using unfiltered speech, and a DNN-LSTM fusion, with the DNN-LSTM[23] serving as the benchmark. The proposed classification model achieves the best accuracy. For CT1 with 5 target classes, the proposed classification model achieves 85.1% accuracy as compared to 75% by the single CNN and 68% by DNN-LSTM Fusion. It is highlighted that the single CNN with global average pooling performs better than the DNN-LSTM fusion.

Classification experiments have been repeated 12 times, and the classification accuracy by the proposed classification model has been shown to be considerably higher than the other two models, in every classification iteration. The Non-parametric Friedman test has been used to analyse the statistical significance of the repeated classification results. The test results in a p-value of 0.0005, which indicates a 0.05% chance of the null hypothesis being true, which confirms that the improvement by the proposed classification model is unlikely to be coincidental. Additional tests have also confirmed that the Mel filter-bank features outperform MFCC, as input features to the proposed model, and CNN fusion using a combination of 4 kHz and 1 kHz for low-pass filtering results in better classification scores as compared to fusion with single filters of 1, 2, or 4 kHz.

Classification Performances for Different Indian Regions

Subsequently, multiple classification metrics have been used to analyse the performance of the proposed model across Indian regions for the different classification tasks. Precision, recall, and F-score, along with confusion matrices have been computed for all the classification tasks.

Recall can be considered as accuracy for each target class and is defined as the ratio of 'true predictions' to 'total test speakers' for a particular target class. On the other hand, precision is defined as the ratio of 'true predictions' to the 'total predictions' for a class, and hence, it penalizes false predictions for each class. Precision is more relevant in certain applications, where false positives are more damaging than false-negative predictions, for instance, in forensic applications. F-score is the harmonic mean for precision and recall ratios, with harmonic mean representing the average of ratios. As such, F-score is an effective classification measure as it accounts for both precision and recall ratios.

The confusion matrix and performance measures for CT1 have been depicted in Figs. 3(a) and 3(b) respectively. The confusion matrix in Fig. 3(a) indicates that most states for CT1 can be identified correctly for the majority of speakers, with the exception of speakers from Kerala, which have been mainly confused as Hindi speakers. 33% of the Kerala speakers have been identified as belonging to Hindi states. For classification scores as can be seen in Fig. 3(b), Kerala is the only state with a low recall score of 62%. The rest of the 4 classes have a recall of around 90%. In terms of precision, the highest precision is obtained by Tamil Nadu with 92%, whilst Kerala has a relatively lower precision score of 75%. Hindi has a comparatively low precision value of 81%, despite its 89% recall value, due to the number of Kerala speakers which have been wrongly predicted as belonging to Hindi states. F-score values balanced precision and recall values, such that the highest F-score value is obtained by Andhra Pradesh with 91%. The lowest F-score value is obtained by Kerala with 68%.

Classification performances regarding Hindi region are evaluated for both English and Hindi speech, using the Hindi-English and Hindi-Hindi datasets respectively, as classification tasks 2–5. Overall classification scores of the proposed classification model along with weighted averages of precision, recall, and F-score, for classification tasks 2–5, are listed in Table 3. The weighted averages across the classes are weighted using total number of test speakers per class for each of the metrics. The model achieves just above chance level accuracy for classification within the Hindi regions. Comparing classification task 2 to classification task 4 and classification task 3 to classification task 5 indicates that all classification scores using English speech are better than those using Hindi speech. Comparison between classification tasks for the same language has been used to evaluate the groupings of Hindi states. Higher average precision values are obtained for

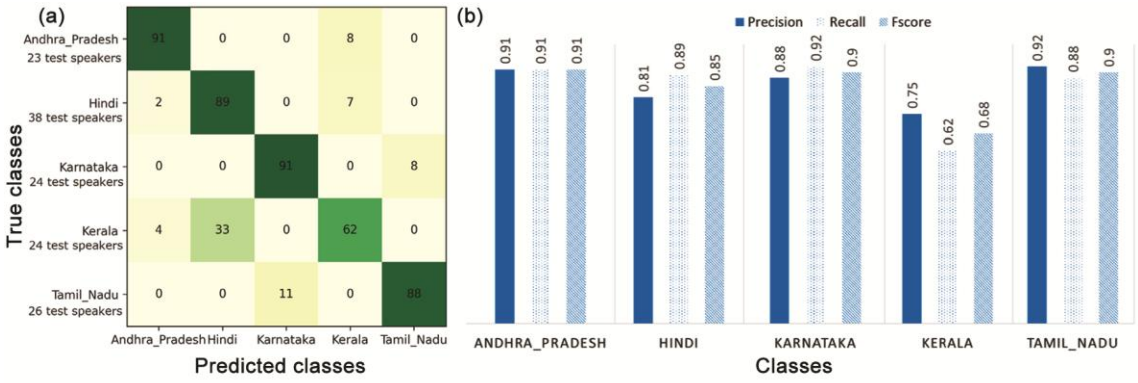

Fig. 3 — Results for Classification Task 1: (a) Confusion matrix, (b) Classification scores

Table 3 — Classification scores for classification tasks 2–5

CT	Target classes	Language	Accuracy	Precision	Recall	F-Score
2	Eastern-Western	English	55.00	0.55	0.55	0.53
3	Northern-Central	English	50.00	0.63	0.49	0.53
4	Eastern-Western	Hindi	50.00	0.49	0.49	0.50
5	Northern-Central	Hindi	42.00	0.56	0.41	0.45

Table 4 — Region-wise classification scores for classifications tasks 2–5

CT	Class	Language	Test speakers	Precision	Recall	Fscore
2	Eastern	English	20	0.56	0.75	0.64
2	Western	English	18	0.55	0.33	0.41
3	Central	English	10	0.29	0.6	0.39
4	Northern	English	28	0.76	0.46	0.58
4	Eastern	Hindi	20	0.52	0.55	0.54
4	Western	Hindi	18	0.47	0.44	0.46
5	Central	Hindi	10	0.23	0.50	0.31
5	Northern	Hindi	28	0.69	0.39	0.50

classification tasks 3 and 5 (to classify between the Northern and Central Hindi states), whilst higher accuracy and average recall values are obtained for classification tasks 2 and 4 (to classify between the Eastern and Western states), using both the Hindi and English languages.

The number of test speakers, precision, recall, and F-score, for each class in classification tasks 2–5 are tabulated in Table 4. It can be seen that precision for classification using the English speech is higher as compared to the Hindi speech for all classes. Similarly, recall for classification using the English speech is higher for all classes, with the exception of Western Hindi states. Results from Tables 3 and 4 indicate that classification using English (L2) performs better than classification using Hindi, the native language (L1) of the speakers.

Discussion

The reported accuracy of up to 85% by the proposed accent classification model, on the Indian-multilingual dataset; significantly out performs the benchmark DNN-LSTM[23] model using the same acoustic features. Moreover, the fusion of models using LPF versions of the signal has been shown to perform considerably better than a single classifier with similar architecture without exploiting LPF. The same trend in results for different L2 English speech datasets confirms the validity of the findings irrespective of the speakers' social backgrounds.

All existing models for accent classification use the entire set of spectral features from speech utterances as a single input for the classifier.[3–6] In contrast, the proposed method utilizes spectral features from multiple copies of speech, each filtered with different cut-off frequencies. This approach effectively narrows down the range of spectral characteristics that the classifier focuses on, making it more specific and targeted. As a result, the classifiers are able to identify accent-discriminating characteristics within specific spectral ranges of speech. The experiments demonstrate that this proposed method, which involves multiple classifiers focusing on specific spectral ranges, particularly in lower frequency bands, outperforms the benchmark techniques discussed in recent literature when tested on the same dataset.

Improvement in accuracy by merging the classification with the low pass filtered input version indicates that convolutional neural networks focusing separately on different levels of information in the frequency domain of speech are effective for accent classification. Moreover, the effectiveness of low pass filtering for classification fusion indicates that most of the accent-discriminating information lies within the lower frequency bands of speech. The significance of lower frequency components for accent classification is encouraging as most of the digital communication media, such as telephone networks, act as low pass filters suppressing the higher frequencies in speech.

The results for classification of speakers within the Hindi regions, show just above chance level performance. The task gets more challenging as the speech considered is read from fixed transcripts and does not contain any dialectical differences. Better results for the same set of speakers in English as compared to Hindi, indicate that classification of accents is relatively easier using L2 speech. Additionally, it is also relatively easier to differentiate the Eastern Hindi speakers from the Western Hindi speakers, and vice versa, as compared to differentiating the Northern Hindi speakers from the Central Hindi speakers, and vice versa.

Since the proposed model is effective in classifying accents based on single sentences, the model may be applicable for forensic applications, where speech samples available as crime-related evidence have commonly short durations. The short duration of the speech necessitates the criminals to be selective of their speech content and avoid dialectical

differences; hence, analysis over speech, read from transcripts may be more applicable to forensic applications.

The proposed model can also be used as a hierarchical classification system. For instance, two of the proposed classification models can be stacked together, using initially the model trained to identify the wider geographical regions and then, specifically for speakers deemed to be from the Hindi region, another trained model for classification tasks 3 or 4, can be used to further identify the speakers from either the Eastern/Western or Northern/Central Hindi regions.

Conclusions

This paper proposed a purely acoustic method for estimating social background of Indian speakers based on their speech accents. The accent classification model is based on the fusion of convolutional neural networks, with global-average pooling as the final layers. Mel Filter-bank features of raw speech signals, along with their LPF copies, are used as input to three similar convolutional neural networks. After training the convolutional networks, their average-pooling layer representations are concatenated and used as input for a Support Vector Machine (SVM) classifier.

The proposed model has achieved 85% accuracy while trying to predict the social background of Indian speakers regarding their native languages. The classification accuracy is 17% higher than the benchmark model implemented for the same dataset. Moreover, results have also indicated that classification of accents is easier using the second language of the speakers, as compared to their native language. Finally, the correlation of classification scores with the speaker origins can be useful for the socio-forensic research community.

It is noted that the findings are valid only for speech, read from transcripts, and do not apply to spontaneous speech. Moreover, the model has also disregarded linguistic features including phonetic, lexical, and syntactic features. Analysis of spontaneous speech, whilst incorporating lexical features for accent classification, represents future research directions of this work.

References

1 Hansen J H L & Hasan T, Speaker recognition by machines and humans: A tutorial review, *IEEE Signal Process Mag*, **32(6)** (2015) 74–99, doi:10.1109/MSP.2015.2462851.

2 Kalluri S B, Vijayasenan D & Ganapathy S, Automatic speaker profiling from short duration speech data, *Speech Commun*, **121(**2019) 16–28, doi:10.1016/j.specom.2020.03.008.

3 Brown G, Automatic accent recognition systems and the effects of data on performance, *Odyssey 2016 Speak Lang Recognit Work*, (2016) 94–100, doi:10.21437/Odyssey.2016-14.

4 Najafian M, Safavi S, Weber P & Russell, M. Identification of British English regional accents using fusion of i-vector and multi-accent phonotactic systems, *Odyssey 2016: Speaker and Language Recognition Workshop*, (2016) 132–139, doi:10.21437/Odyssey.2016-19.

5 Brown G, Exploring forensic accent recognition using the Y-ACCDIST system, *Sixt Annu Conf Int Speech Commun Assoc*, (2016) 305–308.

6 Najafian M, Khurana S, Shan S, Ali A & Glass J, Exploiting Convolutional Neural Networks for Phonotactic Based Dialect Identification, *IEEE Proc Int Conf Acoustics, Speech Signal Process* (IEEE) 2018, 5174–5178, doi:10.1109/ICASSP.2018.8461486.

7 Ferragne E & Pellegrino F, Vowel systems and accent similarity in the British Isles: Exploiting multidimensional acoustic distances in phonetics, *J Phon*, **38(4)** (2010) 526–539, doi:10.1016/j.wocn.2010.07.002.

8 Rajpal A, Patel T B, Sailor H B, Madhavi M C, Patil H A & Fujisaki H, Native language identification using spectral and source-based features, *Proc Interspeech* 2016, 2383–2387, doi:10.21437/Interspeech.2016-1100.

9 Tahir F, Saleem S & Ahmad A, Extracting accent information from Urdu speech for forensic speaker recognition, *Turkish J Electr Eng Comput Sci*, **27(5)** (2019) 3763–3778, doi:10.3906/elk-1812-152.

10 Kalluri S B, Vijayasenan D, Ganapathy S M R R & Krishnan P, NISP: A Multi-lingual Multi-accent Dataset for Speaker Profiling, *IEEE Int Conf Acoustics, Speech Signal Process* (IEEE) 2021, 6953–6957, doi:10.1109/icassp39728.2021.9414349.

11 Ververidis D & Kotropoulos C, Emotional speech recognition: Resources, features, and methods, *Speech Commun*, **48(9)** (2006) 1162–1181, doi:10.1016/j.specom.2006.04.003.

12 Nguyen T H, Chng E S & Li H, Speaker Diarization: An Emerging Research, *Speech and audio processing for coding, enhancement and recognition*, (2014) 229–277, doi:10.1007/978-1-4939-1456-2_8.

13 Snyder D, Garcia-Romero D, Povey D & Khudanpur, S. Deep neural network embeddings for text-independent speaker verification, *in Interspeech*, **2017** (2017) 999–1003, doi:10.21437/Interspeech.2017-620.

14 Snyder D, Garcia-Romero D, Sell G, Povey D & Khudanpur S. X-Vectors: Robust DNN Embeddings for Speaker Recognition, *IEEE Proc Int Conf Acoust Speech Signal Process* (IEEE) 2018, 5329–5333, doi:10.1109/ICASSP.2018.8461375.

15 Goel N K, Sarma M, Kushwah T S, Agrawal D K, Iqbal Z & Chauhan S, Extracting speaker's gender, accent, age and emotional state from speech, *Proc Annu Conf Int Speech Commun Assoc* (*Interspeech*) 2018, 2384–2385, doi:10.21437/Interspeech.2018-3036.

16 Singh R, Raj B & Baker J, Short-term analysis for estimating physical parameters of speakers, *4th Int Conf Biomet Forensic* (IEEE) 2016, 1–6.

17 Ximenes A B, Shaw J A & Carignan C, A comparison of acoustic and articulatory methods for analyzing vowel differences across dialects: Data from American and Australian English, *J AcoustSoc Am*, **142(1)** (2017) 363–377, doi:10.1121/1.4991346.

18 Lundmark M S, Ambrazaitis G & Ewald O, Exploring multidimensionality: Acoustic and articulatory correlates of Swedish word accents, *Proc Annu Conf Int Speech Commun Assoc, in Interspeech* (The International Speech Communication Association) 2017, 3236–3240, doi:10.21437/Interspeech.2017-1502.

19 Weninger F, Sun Y, Park J, Willett D & Zhan P, Deep learning based Mandarin accent identification for accent robust ASR, *Proc Annu Conf Int Speech Commun Assoc, in Interspeech* 2019, 510–514, doi:10.21437/Interspeech. 2019-2737.

20 Brown G, *Segmental content effects on text-dependent automatic accent recognition* (ISCA) 2018, 9–15, doi: 10.21437/odyssey.2018-2.

21 Hughes V & Wormald J, Sharing innovative methods, data and knowledge across sociophonetics and forensic speech science, *Linguist Vanguard*, **6(s1)** (2020) 20180062, doi:10.1515/lingvan-2018-0062.

22 Ge Z, Improved accent classification combining phonetic vowels with acoustic features, *Proc - 2015 8th Int Congress Image Signal Process* (IEEE) (2015) 1204–1209, doi:10.1109/CISP.2015.7408064.

23 Jiao Y, Tu M, Berisha V & Liss J, Accent identification by combining deep neural networks and recurrent neural networks trained on long and short term features, *Proc Annu Conf Int Speech Commun Assoc, in Interspeech* 2016, 2388–2392, doi:10.21437/Interspeech.2016-1148.

24 Soorajkumar R, Girish G N, Ramteke P B, Joshi S S & Koolagudi S G, Text-independent automatic accent identification system for Kannada language, *in Advances in Intelligent Systems and Computing*, **469** (2017) 411–418, doi:10.1007/978-981-10-1678-3_40.

25 Chen J, Huang Q & Wu X, Frequency importance function of the speech intelligibility index for Mandarin Chinese, *Speech Commun*, **83** (2016) 94–103, doi: 10.1016/j.specom.2016.07.009.