Kethireddy, Rashmi; Kadiri, Sudarsana Reddy; Gangashetty, Suryakanth V.

Exploration of temporal dynamics of frequency domain linear prediction cepstral coefficients for dialect classification

Published in:
Applied Acoustics

DOI:
10.1016/j.apacoust.2021.108553

Published: 01/01/2022

Document Version
Publisher's PDF, also known as Version of record

Applied Acoustics 188 (2022) 108553

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Exploration of temporal dynamics of frequency domain linear prediction cepstral coefficients for dialect classification

Rashmi Kethireddy [a,*], Sudarsana Reddy Kadiri [b,*], Suryakanth V. Gangashetty [c]

[a] Speech Processing Laboratory, International Institute of Information Technology, Hyderabad, India
[b] Department of Signal Processing and Acoustics, Aalto University, Finland
[c] Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India

ARTICLE INFO

Article history:
Received 14 February 2021
Received in revised form 29 October 2021
Accepted 23 November 2021

Keywords:
Frequency domain linear prediction
Long temporal variations
Dialect classification
i-vectors
x-vectors

ABSTRACT

Speakers exhibit dialectal traits in speech at sub-segmental, segmental, and supra-segmental levels. Any feature representation for dialect classification should appropriately represent these dialectal traits. Traditional segmental features such as mel-frequency cepstral coefficients (MFCCs) fail to represent sub-segmental and supra-segmental dialectal traits. This study proposes to use frequency domain linear prediction cepstral coefficients (FDLPCCs) for dialect classification inspired by its long temporal summarization during pole estimation. The i-vectors and x-vectors derived from both baseline (MFCCs, linear prediction cepstral coefficients (LPCCs), perceptual LPCCs (PLPCCs), RASTA filtered PLPCCs (PLPCC-R) and proposed (FDLPCC) features are used for identifying the dialects with support vector machine (SVM) and feed-forward neural network (FFNN) as classifiers. Proposed FDLPCC features have shown to perform better than baseline features such as MFCCs and PLPCC-Rs (best among LPCCs variants) by an absolute improvement of 3.4% and 3.9% (in unweighted average recall (UAR)), with i-vector + SVM system and 1.6% and 4.6% (in UAR), i-vector + FFNN system respectively. It is also found that there exists a complementary information between the proposed and baseline features. Furthermore current studies are compared with previous studies and it is found that performances of current studies are better than previous studies.

1. Introduction

Dialect identification/classification refers to a process of automatically identifying the dialect of a speaker, which indirectly provides the regional origin of the speaker. In this study, the dialect classification task is carried out from the acoustic speech signal.

Dialectal variations can be observed at the phonemic level, syllabic level, or sentence level. Phonemic level variations include the variations in the distribution of sounds and variations in articulatory trajectories within the same sound across dialects [1]. Syllabic level variations across dialects occur due to variations in stress patterns, intonation contour, duration, and articulatory trajectories based on the rules defined for respective dialect [1,2]. Sentence level variations across dialects occur due to variations in sentence-level intonation and higher-level linguistic factors such as usage of words, i.e., vocabulary [3]. From the above discussion, it is evident that the dialect discriminant information can be found not only by observing a single sound unit (phonemic/syllabic), but also temporal dynamics across the sound units.

Conventional short-term spectral features such as mel frequency cepstral coefficients (MFCCs) are derived by windowing the signal with a window of length 10–30 ms and incorporate weak temporal context using delta coefficients (Δ, and ΔΔ), and shifted delta coefficients (SDCs) [4,5]. These windowed representations may fail to represent the instantaneous burst representations of stops and fricatives and also may fail to represent temporal dynamics across windows [6,7,8].

Representation for temporal dynamics of the speech signal can be obtained at the acoustic level or at the phonetic level. Acoustic-level temporal dynamics can be represented by segmenting the speech signal into syllables either manually or automatically. In [9], the speech signal is segmented into pseudo-syllables, and the acoustic variations such as pitch, rhythm, and duration are investigated for dialect classification. In [10], supra-segmental prosodic variations obtained from pseudo-syllables are modeled using n-gram language model. To take the advantage of temporal context, two models (stochastic trajectory model (STM) and parametric trajectory model (PTM)) are investigated on segmental cepstral

* Corresponding authors.
E-mail addresses: rashmi.kethireddy@research.iiit.ac.in (R. Kethireddy), sudarsana.kadiri@aalto.fi (S.R. Kadiri).

https://doi.org/10.1016/j.apacoust.2021.108553
0003-682X/

coefficients in [11]. Temporal dynamics can also be modelled using higher linguistic features such as phones [11–18]. The methods in this approach involves a phone recognizer and modelling techniques such as phone recognition followed by language model (PRLM) and parallel-PRLM (PPRLM) [11–15]. These approaches require an external phone recognizer and often the dialect identification accuracy depends on the performance of phone recognizer. To overcome this, we investigate the effectiveness of acoustic features that captures the longer temporal context.

In the present study, the effectiveness of frequency domain linear prediction cepstral coefficients (FDLPCCs) which has the ability to capture the longer temporal context are investigated for dialect classification. Traditional linear prediction, i.e., time-domain linear prediction (TDLP) analysis estimates the spectral peaks by computing auto-correlation of a signal. By duality principle, frequency domain linear prediction (FDLP) estimates temporal peaks by computing auto-correlation of discrete cosine transform (DCT) sequence [6,19–22]. Unlike conventional short-term spectral feature extraction methods, the sub-band FDLP envelope captures extended temporal context as the estimated temporal peaks are the resultant of long-timescale summarization [6]. We hypothesize that the long temporal nature of FDLP spectrum may be advantageous in discriminating dialects.

Fig. 1 illustrates the temporal variations in terms of amplitude envelopes across six sub-bands (i.e., Fig. 1 (b)-(f)) for the word 'adult' spoken by an American speaker (shown in Fig. 1 (i)) and by a British speaker (shown in Fig. 1 (ii)). The speech signals are shown in subplots (a) in Fig. 1. From the figure, it can be clearly observed that the temporal variations in stress patterns between American and British speakers are different. American speaker stressed on the second syllable (see Fig. 1 (i) in time interval of 250 to 500 ms) while the British speaker stressed on the first syllable (see Fig. 1 (ii) in time interval of 100 to 200 ms) of a bi-syllabic word. Inspired by this observation, FDLP based cepstral coefficients are investigated for dialect classification in this study.

The deep neural network (DNN) architectures with convolution neural network (CNN) and time delay neural network (TDNN) models were investigated in the previous studies [22–31] which could capture long temporal context. Therefore, the proposed (FDLPCC) and baseline features are investigated with deep embeddings (x-vectors) derived from TDNN model along with traditional factor analysis based i-vectors. They are also compared to previous studies worked with UT-Podcast using DNN architectures [23].

The contributions of this study are as follows:

- Application of FDLPCCs for dialect classification based on the hypothesis that FDLP captures the longer temporal dynamics.
- Analysis of different temporal context representations such as delta & double delta ($\Delta + \Delta\Delta$), and shifted delta cepstra (SDC) coefficients for baseline and proposed features.
- Investigating the effect of different number of static cepstral coefficients for dialect classification.
- Investigating the effect of different pole orders during the extraction of sub-band FDLP envelopes for dialect classification.
- Investigation of i-vectors and x-vectors derived from baseline and proposed features with support vector machine (SVM) and feed-forward neural network (FFNN) classifiers.
- Comparison of proposed study results with the previous studies that uses UT-Podcast corpus.

The remainder of the paper is organized as follows: Section 2 gives the description of FDLP method to extract the FDLPCCs. Details of the proposed dialect classification system are given in Section 3. Section 4 gives the experimental setup, which includes the corpus used, baseline features used for comparison and evaluation metrics considered. Results of the proposed dialect classification system and baseline systems are discussed in Section 5 along with the comparison to previous studies that uses UT-Podcast corpus. Finally, Section 6 concludes the study.

2. FDLPCCs feature extraction

Frequency domain linear prediction (FDLP) is an efficient method for auto regressive (AR) modelling of temporal envelopes of speech signal [6,19–22]. The AR model approximates the power spectrum of the speech signal in time domain linear prediction (TDLP), whereas in FDLP, an all pole model is fitted to the Hilbert envelope (squared magnitude of the analytic signal). As the estimated temporal peaks are the resultant of longer time signal, they capture finer details of the linguistic units. We hypothesize that the long temporal nature of FDLP spectrum may be advantageous in discriminating dialects. The extraction of frequency domain linear prediction cepstral coefficients (FDLPCCs) from speech signal involves two stages as shown in Fig. 2. The first stage (first seven blocks in the figure) involves the estimation of sub-band temporal envelopes and the second stage (next three blocks in the figure) involves the extraction of cepstral coefficients from sub-band FDLP envelopes. The steps involved in estimation of sub-band FDLP envelopes are described in Section 2.1, and the extraction of cepstral coefficients (i.e., FDLPCCs) from FDLP envelopes are described in Section 2.2.

2.1. FDLP method

This section describes the steps involved in the estimation of sub-band FDLP envelopes from speech signal [20]. They are:

- Speech signal $s[n]$ is pre-emphasized to remove the low frequency variations caused due to recordings, and to emphasize high frequency components.

$$x[n] = s[n] - \alpha s[n-1] \tag{1}$$

- DCT full-band sequence is computed by applying DCT over the pre-emphasized signal ($x[n]$) for every second. Unlike short-time segmental feature extraction methods, spectral transformation is done over a long temporal signal.

$$y[k] = a[k]\sum_{n=0}^{N-1} x[n]cos\left(\frac{(2n+1)\pi k}{2N}\right), \tag{2}$$

where $k = 0, 1, 2 \ldots N-1$ and

$$a[k] = \begin{cases} \frac{1}{\sqrt{N}} & k = 0 \\ \sqrt{\frac{2}{N}} & k = 1, 2, \ldots, N-1 \end{cases}$$

- Sub-band DCT components are derived by windowing the full-band DCT sequence. The sub-band DCT sequence for a band f (critical band windowing) is represented by $\hat{y}[f]$.
- Analogous to TDLP, applying DFT over the squared magnitude of analytic signal gives auto-correlation of spectral coefficients. The inverse DFT (IDFT) of zero-padded DCT sequence is called even symmetric discrete time analytic signal. The analytic signal derived from each sub-band DCT component is given by:

$$q_a[n] = IDFT(\hat{y}[f]) \tag{3}$$

Autocorrelation coefficients for each sub-band spectrum $\hat{y}[f]$ is derived by applying DFT over each sub-band analytic signal, as given by:

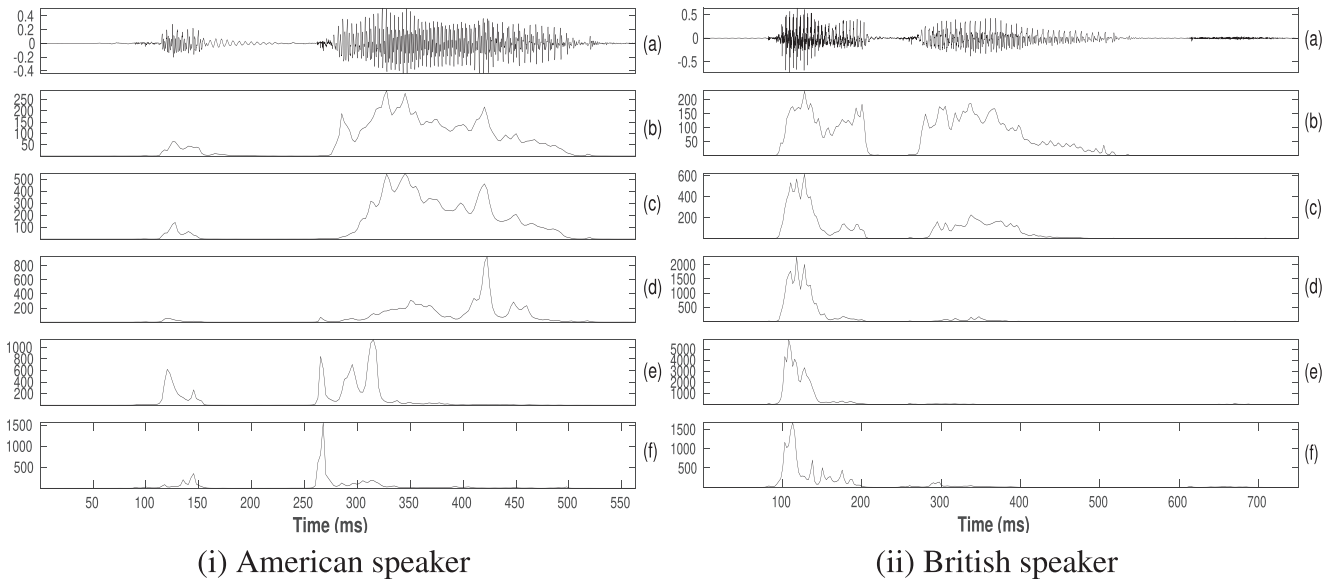

Fig. 1. Illustration of sub-band temporal envelopes estimated using FDLP for the word 'adult' spoken (i) by an American speaker and (ii) by a British speaker.

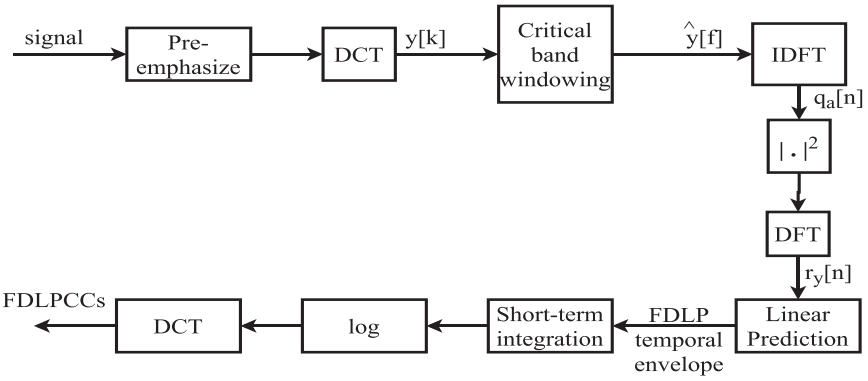

Fig. 2. Block diagram describing the steps involved in extraction of FDLPCCs.

$$r_y[\tau] = DFT(|q_a[n]|^2) \tag{4}$$

- Similar to TDLP, these autocorrelations are used to obtain linear prediction coefficients that are smoothed approximation of sub-band Hilbert envelopes. The LP order (or pole order) to estimate LPCs modulate the efficient representation of sounds. The approximation of sub-band Hilbert envelopes estimated using LPCs is referred as sub-band FDLP envelope in this study. The sub-band FDLP envelope captures extended temporal context as the estimated temporal peaks are the resultant of long-timescale summarization.

2.2. Extraction of FDLPCCs

- Energies in set of sub-band FDLP envelopes are integrated in a long-term analysis window to obtain FDLP short-term frames. To be analogous to short-time segmental feature extraction methods, the window length and window shift are similar to conventional methods.
- DCT is applied over logarithm of integrated FDLP energies across sub-bands within a frame to obtain FDLPCCs for each frame.

3. Dialect classification system

This section describes the stages involved in the proposed dialect classification system. The proposed system consists of three main parts as given in Fig. 3: front-end feature extraction, back-end pre-processing, and classification. The feature extraction part includes the extraction of FDLPCCs (static), and then the computation of temporal context by delta & double delta ($\Delta + \Delta\Delta$) and shifted delta cepstral (SDC) coefficients. Back-end pre-processing involves the extraction of fixed length i-vectors/x-vectors from the variable-length features. The last part classifies the fixed length i-vectors/x-vectors into one of the dialect classes by using support vector machine (SVM) and feed-forward neural network (FFNN) classifiers.

3.1. Parameters used for FDLPCCs extraction:

In this study, the entire signal is considered to obtain the full-band DCT sequence, and then the DCT sequence is multiplied with mel-band Gaussian windows. Typically, the number of mel-band Gaussian windows are given by:

$$n_{mel-bands} = \lceil \mathrm{F}_{\mathrm{hz2mel}}(\frac{fs}{2}) \rceil, \tag{5}$$

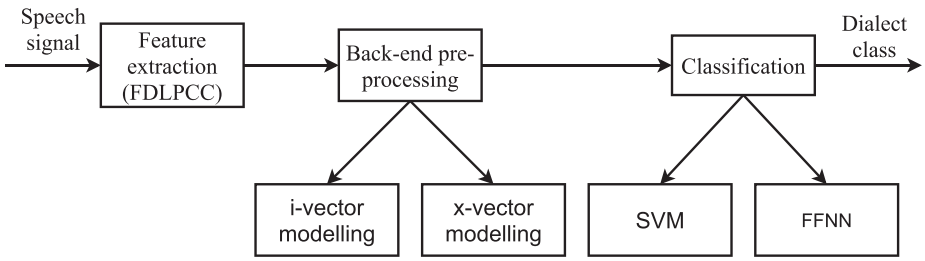

Fig. 3. Block diagram showing the dialect classification system with proposed features (FDLPCC) and back-end pre-processing approaches (i-vector and x-vector modelling).

where *fs* is the sampling frequency in Hertz (Hz) and F_{hz2mel} is a function that converts Hz to mel using Slaney's auditory tool box [32] which will result in 37. However, different number of mel-bands such as 13, 37, 80, 128 and 160 were investigated and it was observed that number of mel-bands such as 37 and 80 gave better performance compared to others. In all the experiments of the study, 37 mel-bands are used.

Autocorrelation formulation of linear prediction is used to estimate temporal poles for each sub-band FDLP envelope. The number of temporal poles is set to 160, similar to previous studies [21]. However, the effect of number of temporal poles is investigated in Section 5.4 for dialect classification. The gain normalized sub-band FDLP temporal envelopes are integrated along time axis within a window of 25 ms and half of it is used as window shift. Static FDLPCCs are obtained by applying DCT over logarithm of integrated FDLP energies across sub-bands within a frame. We investigated the effect of number of static cepstral coefficients (by varying from 13 to 60) on performance of dialect classification. From static coefficients, $\Delta + \Delta\Delta$ and SDC coefficients [5] are also derived, which are also investigated to see their effectiveness on dialect classification [1].

3.2. Back-end pre-processing

Back-end pre-processing involves the extraction of fixed length i-vectors/x-vectors from the variable-length FDLPCCs (based on the number of frames in an utterance).

3.2.1. i-vector extraction

Extraction of i-vectors is motivated by the factor analysis modelling, where features are represented in terms of uncorrelated components [33]. In this, GMM-UBM (trained on all utterances) model is adapted to represent a variable-length utterance in terms of fixed representation called super-vectors. Later by the factor analysis, super-vectors are further compressed to retain only an uncorrelated low-dimensional components of super-vectors, which are called i-vectors. Adapted super-vector $\mathbf{m}$ can be represented as $\mathbf{m} = \mathbf{M} + \mathbf{Tv}$; where $\mathbf{M}$ represent mean super-vector obtained by training GMM-UBM with features from all dialects, $\mathbf{T}$ represents total-variability matrix and $\mathbf{v}$ represents i-vectors. The means and variances of GMM-UBM are initialized using k-means clustering. Initial experiments were conducted by varying number of Gaussian components (256, 512, 640, and 1024) with i-vector system trained with MFCC features. From the experiments, it was observed that 640 Gaussian components performed better than all others and hence the number of Gaussian components is set to 640 across all the experiments. GMM is trained with all the dialects to obtain means of GMM-UBM model (represented by *M*) from the pre-initialized means and variances using k-means clustering. Then the means of GMM-UBM are adapted to each dialect class (represented by *m*). Factor analysis model is trained for 5 epochs to learn the total variability matrix (represented by *T*) using Baum welch statistics. From means (*m* and *M*) and learnt total variability matrix (*T*), 100-dimensional i-vectors are computed for each utterance. More details about i-vector extraction can be found in [33,35]. Matlab toolbox [2] is used for implementing i-vector framework [36].

3.2.2. x-vector extraction

X-vectors were first introduced for extraction of speaker embeddings [37], later extended to other speech applications such as speech and language recognition [24,26,38]. The deep embeddings extracted from the deep neural network (DNN) trained to classify dialects are supposed to contain dialect discriminant information. These embeddings are termed as x-vectors. Traditionally, time-delay neural networks (TDNNs) that are trained with long temporal context are used as DNN architectures to extract x-vectors. The DNN architecture to extract x-vectors contains TDNN layers (TD-layer), fully connected layers (FC-layer), and pooling layer. TD-layer is defined by input dimension, output dimension, and context, and FC-layer is defined by input and output dimensions. Fig. 4 shows the block diagram of TDNN architecture to extract x-vectors. TDNN is trained by the baseline and proposed features (static cepstral coefficients) to classify dialect, and x-vectors (or deep dialect embeddings) are extracted from FC-7 layer. TDNN is trained for 50 epochs with optimizer as adamW (adam with weight decay). The dimension of x-vectors is 512 and ReLU activation is applied across all the layers. Overall temporal context captured by the TDNN in Fig. 4 to extract x-vectors is 23 frames. The configurations and architecture of TDNN are similar to [38,40]. X-vector framework is developed using kaldi[3] with PyTorch libraries [38,40].

3.3. Classification

Finally, classifier predicts the dialect class using support vector machine (SVM)/feed-forward neural network (FFNN) classifiers. In the experiments, about 65% of the data is used for training and 35% of data is used for testing as in [41]. A random split 25% of train data is used as validation data. SVM classifier [4] of this study is configured with a linear kernel. The regularization parameter C of SVM [42] is set to value between 0.1 to 1.0 during validation.

FFNN classifier [5] of this study is a four layered neural network with 2 hidden layers with dimension of 64. ReLu is used as activation function at each layer. Training is configured by a learning rate of 0.01 with stochastic gradient descent (SGD) as optimizer and the

[1] https://github.com/iiscleap/FeatureExtractionUsingFDLP.

[2] https://github.com/wangwei2009/MSR-Identity-Toolkit-v1.0

[3] https://github.com/Snowdar/asv-subtools

[4] https://scikit-learn.org/stable/modules/svm.html

[5] https://github.com/yunjey/pytorch-tutorial/tree/master/tutorials/01-basics/feedforward_neural_network

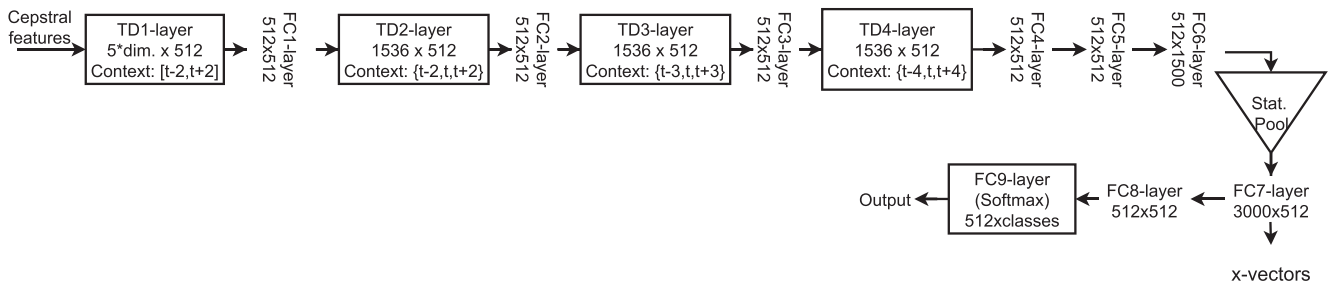

Fig. 4. Block diagram showing the architecture of time-delay neural network (TDNN) that is used to extract x-vectors (deep dialect embeddings).

batch size is set to 16. Maximum number of epochs is set to 2000 and the model with higher UAR on validation data is considered for testing. For stable initialization of weights of neural network, manual seed is set to 1024.

4. Experimental setup

This section gives the details of the corpus, evaluation metrics considered, and baseline features used for comparison along with the configurations used in extraction of features.

4.1. Corpus

The study uses the spontaneous speech corpus which was collected by crawling web-based podcasts, which mainly contain interviews and it is named as UT-Podcast [41]. The corpus consists of three broad dialects of English: AU (Australian), UK (Britain), and US (American). Since the data is spontaneous, it has many dialectal variations. The total duration of speech used in train set is 5.2 hrs with 2.1 hrs of AU, 1.2 hrs of UK, and 1.9 hrs of US. The total duration of speech used in the test set is 3.2 hrs with 1.6 hrs of AU, 0.4 hrs of UK, and 1.2 hrs of US. All the speech utterances are pre-processed to remove non-speech segments and the resultant utterance average length is of about 17 *sec*. The sampling frequency of the data is 8 kHz.

4.2. Baseline features for comparison

The most popular and conventional Mel-frequency cepstral coefficients (MFCCs) [43] and three variants of linear prediction based features (linear prediction cepstral coefficients (LPCCs) [44], Perceptual LPCCs (PLPCCs) [45], RASTA filtered perceptual LPCCs (PLPCC-R) [46]) are considered as baseline features for dialect classification. The baseline system architecture and configuration are similar to that of the proposed system. For all the features extraction, a window size of 25 ms and half of the window length are considered as window shift. Autocorrelation formulation is used in all three variants of linear prediction based features. Both the baseline and proposed feature representations are investigated by varying number of static cepstral coefficients (by varying from 13 to 60). From static coefficients, $\Delta, \Delta\Delta$, and SDC coefficients [5] are also derived, which are also investigated to see their effectiveness on dialect classification. For Δ and $\Delta\Delta$ computation, a context of three is considered, and for SDCs a standard configuration of N-d-p-K (N-1-3-7) is considered, where N denotes the dimension of the static cepstral coefficients, d denotes the delay/advance from the present frame; p is the shift between consecutive delta computations; and K such delta computations are concatenated to form N×K-dimensional SDC coefficients.

4.3. Evaluation metric

Unweighted average recall (UAR) is considered as the evaluation metric to assess the baseline and proposed dialect classification systems. Further, we also report class-wise accuracies for the best-configured baseline and proposed systems.

5. Results and discussion

In this section, both i-vector and x-vector representations derived from baseline (MFCC, LPCC, PLPCC, and PLPCC-R) and proposed (FDLPCC) frame-level features are investigated (in Section 5.1) for dialect classification. To find the best configurations, i-vector representations derived from baseline and proposed features are investigated with two temporal contexts (i.e., static+Δ +$\Delta\Delta$ and static + SDCs) and by varying static cepstral orders from 13 to 60 (13, 20, 30, 40, 50, and 60). The higher number static coefficients are included to determine whether higher order coefficients contain any additional information useful for the classification of dialects. X-vector representations derived from 40-dimensional static cepstral coefficients (for baseline and proposed) are investigated for dialect classification, which are shown to perform better with i-vector representations.

The existence of complementary information is investigated in Section 5.2 by fusing at frame- level (F-level) and utterance-level (U-level) of baseline and proposed features for both i-vector and x-vector approaches. The results of the present studies with SVM trained with i-vectors/x-ve- ctors and FFNN trained with i-vectors/x-vectors (derived from baseline and proposed features) are compared with the previous studies (in Section 5.3) with both conventional and modern deep neural network (DNN) classifiers. Further, the effect of the number of temporal poles in the FDLPCC feature extraction is investigated for dialect classification in Section 5.4.

5.1. Effect of cepstral order and temporal context

Table 1 shows the performances for the baseline and proposed features with static cepstral coefficients, by varying number of static cepstral coefficients from 13 to 60 (13, 20, 30, 40, 50, and 60). From the table, it can be observed that among the baseline features, MFCC features performed better using 40-dimensional static cepstral coefficients. LPCCs and PLPCCs performed better using 60-dimensional static cepstral coefficients. PLPCC-R found to be better with 30-dimensional static cepstral coefficients, while FDLPCCs are better at 20-dimensional. Among all the features, proposed FDLPCCs gave best performance (77.3%) and the performance of FDLPCCs is consistently better at lower dimensional static coefficients (10, 20 and 30) compared to all the baseline features. Also, it can be observed that the number of static cepstral coefficients

Table 1
Performances (in UAR %) for the baseline and proposed features with static cepstral coefficients, by varying cepstral coefficients dimension from 13 to 60 (13, 20, 30, 40, 50, and 60) (i-vector approach).

Features/ #static coeff.	static cepstral coefficients					
	13	**20**	**30**	**40**	**50**	**60**
MFCC	69.6	73.1	72.6	**75.6**	73.9	71.1
LPCC	69.6	67.3	67.5	66.9	69.4	**69.8**
PLPCC	67.4	69.2	71.4	70.1	68.9	**72.1**
PLPCC-R	66.7	69.4	**72.2**	72.0	70.9	70.0
FDLPCC	71.8	**77.3**	76.2	68.1	67.8	66.2

to be considered are not unique for various feature representations.

Table 2 shows the performances for the baseline and the proposed features with static+Δ+ΔΔ (e.g., results in 39-dimension for 13-dimension static coefficients), by varying number of static cepstral coefficients from 13 to 60. From the table, it can be observed that among the baseline features, MFCC features performed better with 20-dimension (i.e., 60-dimension for static+Δ+ΔΔ). LPCCs performed better at 50-dimension (i.e., 150-dimension for static+Δ+ΔΔ), and PLPCCs performed better with 60-dimension (i.e., 180-dimension for static+Δ+ΔΔ). PLPCC-R features found to be better with 20-dimension (i.e., 60-dimension for static+Δ+ΔΔ) while FDLPCCs is better at 13-dimension (i.e., 39-dimension for static+Δ+ΔΔ). Among all the features, proposed FDLPCCs gave best performance (81.3%) and the performance of FDLPCCs is consistently better at lower dimensional static coefficients (10, 20, 30 and 40) compared to all the baseline features.

Table 3 shows the performances for the baseline and the proposed features with static + SDCs (e.g., results in 13+(7×13)= 104-dimension for 13-dimension static coefficients), by varying number of static cepstral coefficients from 13 to 60. From the table, it can be observed that among the baseline features, MFCC features performed better with 20-dimension (i.e., 20+(7×20)=160- dimension for static + SDCs). LPCC features performed better with 40 (i.e., 40+(7×40)=320- dimension for static + SDCs). On the other-hand, PLPCC and PLPCC-R features performed better with 30 (i.e., 30 +(7×30)=240-dimension for static + SDCs). Among all the features, proposed FDLPCCs gave best performance (79%) at 30 (i.e., 30 +(7×30)=240-dimension for static + SDCs) and the performance of FDLPCCs is consistently better at lower dimensional static coefficients(10, 20 and 30) compared to the baseline features.

From the Tables 1–3, it can be observed that proposed FDLPCCs shown better performance in comparison to the baseline features. Also, it can be observed that most of the features performed better for static+Δ+ΔΔ. UAR and class-wise accuracies are given in Table 4 for the best configurations (Tables 1–3) of baseline and proposed features. From the results, it can be observed that all the features are more accurate in detecting AU and US dialects. On the other-hand, the proposed FDLPCCs shown significantly better discrimination of all the classes including the UK dialect even though the class strength is low.

From Table 1, it can be observed that all the features performed reasonably well with 40- dimensional static cepstral coefficients. Hence 40-dimensional cepstral coefficients are used to train TDNN for extracting x-vectors. Table 5 shows the performances (in UAR and class-wise accuracies) of x-vector approach for both baseline and proposed features. All the features of x-vector approach performed better when compared to static cepstral coefficients of i-vector approach (See Table 1). In comparison to best configured i-vector approach (as in Table 4), performance of x-vector approach for all features is inferior to i-vector approach (except for MFCC). This inferior performance of x-vector approach may be due to imbalanced classes and sparsity of UT-corpus.

5.2. Existence of complementary information

To know the existence of complementary information between baseline and proposed features, experiments are carried out by fusing at frame level (F-level) and utterance level (U-level) for both the modelling approaches (i-vectors and x-vectors). In F-level fusion of i-vector approach, 100-dimensional i-vectors are extracted from fused baseline and proposed features (static+Δ+ΔΔ). In U-level fusion of i-vector approach, 100-dimensional i-vectors are extracted from each of baseline and proposed features, resulted in 200-dimensional i-vectors. In F-level fusion of x-vector approach, 512-dimensional x-vectors are extracted from fused baseline and proposed features (static). In U-level fusion of x-vector approach, 512-dimensional x-vectors are extracted from each of baseline and proposed features, resulted in 1024-dimensional x-vectors. Table 6 shows the performances (in UAR %) of fusion experiments (column 5 and 6) along with individual feature performances (column 3 and 4) for dialect classification. From Table 6 with i-vector approach, it can be observed that U-level fusion showed higher complementary information compared to F-level fusion (except for fusion of MFCC and FDLPCC). With x-vectors approach, it can be observed that both F-level and U-level fusion shown higher complementary information in all the cases in comparison to individual features. Between i-vector approach and x-vector approach, i-vector approach seems to be better for all the fusion sets in both F-level and U-level. This is due to the inferior performance obtained with x-vectors of individual features, as discussed in Section 5.1. Overall, these results indicates the existence of complementary information between the baseline short-term features and proposed long-term FDLP features.

5.3. Comparison of current studies with previous studies

This section compares the results obtained in the current study (i-vectors/x-vectors derived baseline and proposed features with SVM and FFNN classifiers) with the previous studies [40,23]. In [41], both text based and audio based approaches were investigated. In text based approach, term-frequency and inverse document frequency (TF-IDF) was exploited. TF-IDF measures the originality of word in a document. In audio based approach, GMM super-vectors and i-vectors were used with SVM classifier. A fusion of both text and audio approach is also investigated. In [23], DNN classifiers such as feed-forward neural network (FFNN), five-layer convolution neural network (CNN), AlexNet, VGG-11, and ResNet-18 trained with STFT-spectrogram are investigated. In this study, corpus is modified by segmenting the utterances of UK dialect to handle imbalanced classes. FFNN is a DNN classifier with three fully connected layers and five-layer CNN is a DNN classifier with five convolution layers for segmental-level processing and fully connected layers for utterance-level processing. The other DNN classifiers, AlexNet [47], VGG-11 [48], and ResNet-18 [49] are

Table 2
Performances (in UAR %) for the baseline and proposed features with static+Δ+ΔΔ, by varying static cepstral coefficients dimension from 13 to 60 (i-vector approach).

Feature/ #static coeff.	static +Δ+ΔΔ					
	13	20	30	40	50	60
MFCC	74.5	**77.2**	75.0	73.0	73.4	74.2
LPCC	67.3	69.3	68.6	71.5	**74.4**	73.4
PLPCC	68.7	70.5	**75.4**	72.1	71.6	71.6
PLPCC-R	75.3	**76.6**	74.9	73.1	73.5	71.2
FDLPCC	**81.3**	79.3	78.5	75.9	67.6	67.6

Table 3
Performances (in UAR %) for the baseline and proposed features with static + shifted delta cepstra (SDC), by varying static cepstral coefficients dimension from 13 to 60 (i-vector approach).

Features/ #static coeff.	static + SDCs					
	13	20	30	40	50	60
MFCC	76.8	**77.9**	77.1	76.1	73.9	68.8
LPCC	68.0	69.4	67.2	**71.2**	70.6	66.9
PLPCC	70.0	73.3	**74.1**	73.5	71.8	74.1
PLPCC-R	75.5	75.6	**77.4**	75.0	75.9	75.0
FDLPCC	78.2	78.4	**79.0**	72.1	69.0	64.7

Table 4
Performances (in UAR% and class-wise accuracies) for the baseline and proposed (FDLPCCS) features with the best configurations (from Tables 1–3) (i-vector approach).

Features/Class	UAR	AU	UK	US
MFCC (static + SDC)	**77.9**	87.3	56.1	90.4
LPCC (static +Δ+ΔΔ)	**74.4**	**88.8**	46.0	88.3
PLPCC (static +Δ+ΔΔ)	**75.4**	86.7	59.5	80
PLPCC-R (static + SDC)	**77.4**	84.0	62.9	85.4
FDLPCC (static +Δ+ΔΔ)	**81.3**	**86.1**	**66.3**	**91.6**

Table 5
Performances (in UAR% and class-wise accuracies) for the baseline and proposed (FDLPCCS) features with x-vector approach.

Features/Class	UAR	AU	UK	US
MFCC	**76.7**	88.9	56.2	85.0
LPCC	**73.4**	85.8	52.8	81.7
PLPCC	**73.1**	81.3	53.9	84.2
PLPCC-R	**74.4**	78.0	65.2	80.0
FDLPCC	**75.4**	67.4	67.4	81.3

Table 6
Performances (in UAR %) obtained for fusion of baseline and proposed features both with i-vector and x-vector approaches at frame (F-level) and utterance levels (U-level).

Approach	Fusion of feats	Feat1	Feat2	Fusion	
	(Feat1 + Feat2)	UAR	UAR	F-level	U-level
i-vectors	MFCC + FDLPCC	74.5	81.3	84.0	83.2
	LPCC + FDLPCC	67.3		78.0	81.4
	PLPCC + FDLPCC	68.7		79.2	83.3
	PLPCC-R + FDLPCC	75.3		78.5	83.1
x-vectors	MFCC + FDLPCC	76.7	75.5	80.9	76.6
	LPCC + FDLPCC	73.4		81.8	80.3
	PLPCC + FDLPCC	73.1		76.4	80.0
	PLPCC-R + FDLPCC	75.4		76.4	76.7

typical deep architectures with varied number of convolution layers.

Table 7 shows the performance of dialect classification (in UAR % and class-wise accuracies) for previous studies and current studies. From the first set of previous studies (rows 3–6) shown in Table 7, it can be observed that audio based approaches performed better than text based approach. Within the audio based approaches, i-vector approach performed better than GMM approach. The fusion of both audio (i-vectors) and text based systems have shown an improvement in performance by 2.4% relative UAR than i-vector system alone. It can be observed that the current study with MFCC, PLPCC-R, and FDLPCC of i-vector + SVM approach

Table 7
Comparison of current studies with previous dialect classification models over UT-Podcast corpus (in UAR% and class-wise accuracies).

Arch. type	**UAR**	AU	UK	US
Text and audio based approaches from previous studies [40]				
Audio System (GMM)	**60.3**	85.5	32.6	62.9
Audio System (i-vector)	**74.5**	78.0	61.8	83.8
Text System (TF-IDF logistic regression)	**58.7**	83.1	32.6	60.4
Audio-Text system (Fusion)	**76.3**	86.1	60.7	82.1
DNN classifier from previous studies [22]				
FFNN	**61.4**	70.8	50.6	62.9
Five-layer CNN	**62.8**	64.8	41.6	82.0
AlexNet	**64.9**	58.4	64.0	74.2
VGG-11	**54.4**	55.7	48.3	59.2
ResNet-18	**61.7**	69.3	38.2	77.5
SVM trained with i-vectors (current study)				
MFCC (baseline)	**77.9**	87.3	56.1	90.4
LPCC (baseline)	**74.4**	88.8	46.0	88.3
PLPCC (baseline)	**75.4**	86.7	59.5	80
PLPCC-R (baseline)	**77.4**	84.0	62.9	85.4
FDLPCC (proposed)	**81.3**	86.1	66.3	91.6
FFNN trained with i-vectors (current study)				
MFCC (baseline)	**81.9**	84.6	73.0	87.9
LPCC (baseline)	**74.0**	77.4	62.9	81.7
PLPCC (baseline)	**74.3**	68.1	73.0	81.7
PLPCC-R (baseline)	**78.9**	78.6	68.5	89.6
FDLPCC (proposed)	**83.5**	88.6	78.7	83.8
SVM trained with x-vectors (current study)				
MFCC (baseline)	**76.7**	88.9	56.2	85.0
LPCC (baseline)	**73.4**	85.8	52.8	81.7
PLPCC (baseline)	**73.1**	81.3	53.9	84.2
PLPCC-R (baseline)	**74.4**	78.0	65.2	80.0
FDLPCC (proposed)	**75.4**	67.4	67.4	81.3
FFNN trained with x-vectors (current study)				
MFCC (baseline)	**74.0**	79.5	56.2	86.3
LPCC (baseline)	**68.7**	81.9	48.3	75.8
PLPCC (baseline)	**71.7**	76.2	61.8	76.3
PLPCC-R (baseline)	**71.4**	76.2	61.8	76.3
FDLPCC (proposed)	**70.3**	75.9	59.6	75.4

outperformed the fusion system of previous study by 2.1%, 1.4%, and 6.6% (relative UAR) respectively. Among the second set of previous studies (rows 8–12) shown in Table 7 with DNN classifiers [23], it can be observed that AlexNet classified dialects better than other DNN classifiers. Further, it can also be observed that all the current studies (both i-vectors and x-vectors) with conventional SVM and FFNN classifiers trained with baseline and proposed features performed better than the DNN approaches. The inferior performance of DNN can be attributed to data sparsity of UT-Podcast dialect corpora.

The conventional SVM classifier with i-vectors modelled from MFCC, LPCC, PLPCC, PLPC- C-R, and FDLPCC outperformed the best DNN classifier (AlexNet) from previous studies by 20.0%, 14.6%, 16.2%, 19.3%, 25.3% (relative UAR) respectively. FFNN classifier with i-vectors modelled from MFCC, LPCC, PLPCC, PLPCC-R, and FDLPCC outperformed the best DNN classifier (AlexNet) from previous studies by 26.2%, 14.0%, 14.5%, 21.6%, and 28.7% (relative UAR) respectively. The conventional SVM classifier with x-vectors modelled from MFCC, LPCC, PLPCC, PLPCC-R, and FDLPCC outperformed the best DNN classifier (AlexNet) from previous studies by 18.2%, 13.1%, 12.6%, 14.6%, and 16.2% (relative UAR) respectively. FFNN classifier with x-vectors modelled from MFCC, LPCC, PLPCC, PLPCC-R, and FDLPCC outperformed the best DNN classifier (AlexNet) from previous studies by 14.0%, 5.9%, 10.5%, 10.0%, 8.3% (relative UAR) respectively.

From the comparison between SVM and FFNN trained with i-vectors, it can be observed that performance of FFNN classifier is significantly better for MFCC, PLPCC-R and FDLPCC features, and equally well performance for others. Unlike to above observation, SVM trained with x-vectors performed better than FFNN with x-vectors for all of the features. From the comparisons between i-vectors and x-vectors derived from baseline and proposed features, it can be observed that i-vectors performed better in all the cases (except for MFCC). The best performance is achieved using i-vectors + SVM and i-vectors + FFNN derived FDLPCCs with 81.3% and 83.5% UAR respectively). From these results, it can be concluded that the long temporal dependencies captured in FDLPCCs are more advantageous for dialect classification, especially for small corpora like UT-Podcast.

5.4. Effect of pole order used in FDLP for dialect classification

Number of poles (pole order) used for FDLPCC extraction, modulate the effective representation of transient sounds (with higher pole order) versus slowly varying sounds (with lower pole order) [6]. This section investigated the effect of pole order (from 13 to 300) used in extraction of sub-band FDLP envelope for dialect classification. Fig. 5 shows the performance (in UAR %) of proposed dialect classification system (along y-axis) for different pole orders (along x- axis) used in extraction of FDLPCCs. From the UAR plot, it can be observed that performance is stable for poles above 50 and the best dialect discrimination can be achieved by using 200 poles. From this, it can be concluded that the high transient sounds

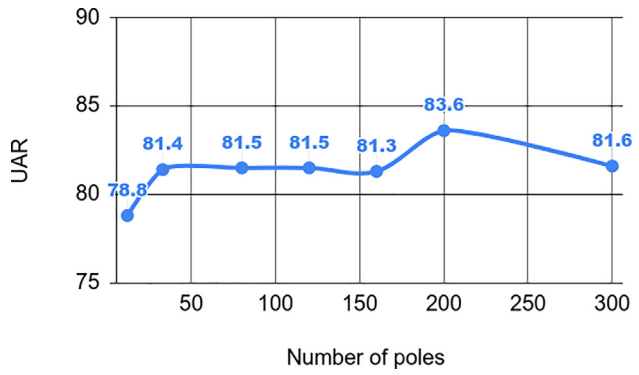

Fig. 5. Performance (in UAR %) by varying number of poles in extraction of FDLPCCs for dialect classification.

majorly contributed in discrimination of major dialects of English (AU, UK, and US).

6. Conclusion

In this study, we proposed to use FDLPCCs for dialect classification which has the potential to capture longer temporal context. From the experiments, SVM trained with i-vectors derived from FDLPCC features were found to perform better than baseline features, such as MFCCs, PLPCCs, RASTA filtered PLPCCs (PLPCC-R), and LPCCs by an absolute improvement of 3.4%, 3.9%, 5.9%, and 6.9% (in UAR), respectively. FFNN trained with i-vectors derived from FDLPCC features were found to perform better than baseline features, such as MFCCs, RASTA filtered PLPCCs (PLPCC-R), PLPCCs, and LPCCs by an absolute improvement of 1.6%, 4.6%, 9.2%, and 9.5% (in UAR), respectively. It was also found that there exists a complementary information between the proposed FDLPCCs and baseline features such as MFCCs, PLPCCs and PLPCC-R (except for LPCCs features). The number of poles modulate the representation of fast varying *vs* slow varying sounds. Investigating different pole orders, it is found that the sub-band FDLP envelope estimated with 200 poles can represent dialect discriminant sounds better. Further from comparison of different modelling approaches, i.e., SVM and FFNN trained with i-vectors and x-vectors (which were derived from baseline and proposed features), it was found that the FFNN trained with i-vectors derived from FDLPCCs performed better than others. SVM trained with i-vectors derived from MFCC, LPCC, PLPCC, PLPCC-R, and FDLPCC, outperformed the best DNN based approach (AlexNet) from previous studies by 19.0%, 14.6%, 16.2%, 18.0%, and 25.3% (in relative UAR) respectively. FFNN trained with i-vectors derived from MFCC, LPCC, PLPCC, PLPCC-R, and FDLPCC outperformed the best DNN based approach (AlexNet) from previous studies by 26.2%, 14.0%, 14.5%, 21.6%, and 28.7% (in relative UAR) respectively. From the experiments in this study, we conclude that long temporal representations (FDLPCCs) help in better dialect discrimination.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The first author would like to thank the University Grants Commission India (Project No. 3582/(NET-NOV2017)) for supporting her PhD. The second author would like to thank the Academy of Finland (Projects 313390 and 330139) for supporting his stay in Finland as a Research Fellow.

References

[1] Mielke J, Carignan C, Thomas ER. The articulatory dynamics of pre-velar and pre-nasal/æ/-raising in English: An ultrasound study. J Acoust Soc Am 2017;142(1):332–49.
[2] Fox R, Jacewicz E. Cross-dialectal variation in formant dynamics of American English vowels. J Acoust Soc Am 2009;126:2603–18.
[3] Huang Y, Guo D, Kasakoff A, Grieve J. Understanding U.S. regional linguistic variation with twitter data analysis. Comput Environ Urban Syst 2016;59:244–55.
[4] Hanani A, Russell MJ, Carey MJ. Human and computer recognition of regional accents and ethnic groups from British English speech. Comput Speech Language 2013;27(1):59–74.
[5] Torres-Carrasquillo PA, Singer E, Kohler MA, Greene RJ, Reynolds DA, Deller Jr JR. Approaches to language identification using Gaussian mixture models and shifted delta cepstral features. Proc Interspeech 2002.
[6] M. Athineos, D.P.W. Ellis, Frequency-domain linear prediction for temporal features, in: Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, 2003, pp. 261–266..
[7] M. Athineos, D.P.W. Ellis, Sound texture modelling with linear prediction in both time and frequency domains, in: Proc. Int. Conf. Acoustics Speech and Signal Processing (ICASSP), Vol. 5, 2003, pp. V–648..
[8] Kethireddy Rashmi, Kadiri Sudarsana Reddy, Alku Paavo, V. Gangashetty Suryakath. Mel-Weighted Single Frequency Filtering Spectrogram for Dialect Identification. IEEE Access 2020;8:174871–9. https://doi.org/10.1109/ACCESS.2020.3020506.
[9] Biadsy F, Hirschberg J. Using prosody and phonotactics in Arabic dialect identification. Proc Interspeech 2009.
[10] Rouas J-L. Automatic prosodic variations modeling for language and dialect discrimination. IEEE Trans Audio Speech Language Process 2007;15(6):1904–11.
[11] Angkititrakul P, Hansen JH. Advances in phone-based modeling for automatic accent classification. IEEE Trans Audio Speech Language Process 2006;14(2):634–46.
[12] Najafian M, Safavi S, Weber P, Russell MJ. Identification of British English regional accents using fusion of i-vector and multi-accent phonotactic system. Proc Odyssey 2016:132–9.
[13] F. Biadsy, J. Hirschberg, N. Habash, Spoken Arabic dialect identification using phonotactic modeling, in: Proc. Workshop on Computational Approaches to Semitic Languages, 2009, pp. 53–61..
[14] M.A. Zissman, T.P. Gleason, D. Rekart, B.L. Losiewicz, Automatic dialect identification of extemporaneous conversational, Latin American Spanish speech, in: Proc. Int. Conf. Acoustics Speech and Signal Processing (ICASSP), Vol. 2, 1996, pp. 777–780..
[15] Richardson FS, Campbell WM, Torres-Carrasquillo PA. Discriminative n-gram selection for dialect recognition. Proc Interspeech 2009:192–5.
[16] N.F. Chen, W. Shen, J.P. Campbell, P.A. Torres-Carrasquillo, Informative dialect recognition using context-dependent pronunciation modeling, in: Proc. Int. Conf. Acoustics Speech and Signal Processing (ICASSP), 2011, pp. 4396–4399..
[17] Chen NF, Tam SW, Shen W, Campbell JP. Characterizing phonetic transformations and acoustic differences across English dialects. IEEE Trans Audio Speech Language Process 2014;22(1):110–24.
[18] Huang R, Hansen JHL, Angkititrakul P. Dialect/accent classification using unrestricted audio. IEEE Trans Audio Speech Language Process 2007;15(2):453–64.
[19] Ganapathy S, Hermansky H. Temporal resolution analysis in frequency domain linear prediction. J Acoust Soc Am 2012;132(5). EL436–EL442.
[20] Ganapathy S, Thomas S, Hermansky H. Temporal envelope compensation for robust phoneme recognition using modulation spectrum. J Acoust Soc Am 2010;128(6):3769–80.
[21] Wickramasinghe B, Irtza S, Ambikairajah E, Epps J. Frequency domain linear prediction features for replay spoofing attack detection. Proc Interspeech 2018:661–5.
[22] Fernando S, Sethu V, Ambikairajah E. Sub-band envelope features using frequency domain linear prediction for short duration language identification. Proc Interspeech 2018:1818–22.
[23] Wu Y, Mao H, Yi Z. Audio classification using attention-augmented convolutional neural network. Knowl-Based Syst 2018;161:90–100.
[24] Dubagunta SP, Magimai-Doss M. Using speech production knowledge for raw waveform modelling based styrian dialect identification. Proc Interspeech 2019:2383–7.
[25] Hanani A, Naser R. Spoken Arabic dialect recognition using x-vectors. Natural Language Eng 2020;26(6):691–700.
[26] Jain A, Upreti M, Jyothi P. Improved accented speech recognition using accent embeddings and multi-task learning. Proc Interspeech 2018:2454–8.
[27] A. Das, K. Kumar, J. Wu, Multi-dialect speech recognition in english using attention on ensemble of experts, in: Proc. Int. Conf. Acoustics Speech and Signal Processing (ICASSP), 2021, pp. 6244–6248..
[28] Zhang Q, Hansen JH. Language/dialect recognition based on unsupervised deep learning. IEEE Trans Audio Speech Language Process 2018;26(5):873–82.
[29] Shon S, Ali A, Glass J. Convolutional neural network and language embeddings for end-to-end dialect recognition. Proc ODYSSEY 2018:98–104.

[30] R. Kethireddy, S.R. Kadiri, S.V. Gangashetty, Learning filterbanks from raw waveform for accent classification, in: Proc. Int. Joint Conf. Neural Networks, 2020, pp. 1–6..
[31] Q. Gao, H. Wu, Y. Sun, Y. Duan, An end-to-end speech accent recognition method based on hybrid CTC/attention transformer ASR, in: Proc. Int. Conf. Acoustics Speech and Signal Processing (ICASSP), 2021, pp. 7253–7257..
[32] Slaney M. Auditory toolbox, Interval Research Corporation. Tech Rep 1998;10:1194.
[33] Dehak N, Kenny P, Dehak R, Dumouchel P, Ouellet P. Front-end factor analysis for speaker verification. IEEE Trans Audio Speech Language Process 2011;19 (4):788–98.
[34] Kenny P, Ouellet P, Dehak N, Gupta V, Dumouchel P. A study of interspeaker variability in speaker verification. IEEE Trans Audio Speech Language Process 2008;16(5):980–8.
[35] H. Lei, Joint factor analysis (JFA) and i-vector tutorial, ICSI. Web. 02 Oct (2011)..
[36] S.O. Sadjadi, M. Slaney, L.P. Heck, MSR identity toolbox v1.0: A matlab toolbox for speaker recognition research, 2013..
[37] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, S. Khudanpur, Deep neural network-based speaker embeddings for end-to-end speaker verification, in: Proc. IEEE Spoken Language Technology Workshop (SLT), 2016, pp. 165–170..
[38] Snyder D, Garcia-Romero D, McCree A, Sell G, Povey D, Khudanpur S. Spoken language recognition using x-vectors. Proc Odyssey 2018:105–11.
[39] Li Z, Zhao M, Hong Q, Li L, Tang Z, Wang D, Song L, Yang C. Ap20-olr challenge: Three tasks and their baselines. In: Proc Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). p. 550–5.
[40] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, S. Khudanpur, Speaker recognition for multi-speaker conversations using x-vectors, in: Proc. Int. Conf. Acoustics Speech and Signal Processing (ICASSP), 2019, pp. 5796–5800..
[41] Hansen JH, Liu G. Unsupervised accent classification for deep data fusion of accent and language information. Speech Commun 2016;78:19–33.
[42] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: Machine learning in python. J Mach Learn Res 2011;12:2825–30.
[43] Davis SB, Mermelstein P. Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Acoust Speech Signal Process 1980;28(4):357–66.
[44] Makhoul J. Linear prediction: A tutorial review. Proc IEEE 1975;63:561–80.
[45] Hermansky H. Perceptual linear predictive (PLP) analysis of speech. J Acoust Soc Am 1990;87(4):1738–52.
[46] Hermansky H, Morgan N. RASTA processing of speech. IEEE Trans Speech Audio Process 1994;2(4):578–89.
[47] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Adv Neural Inform Process Syst 2012;25:1097–105.
[48] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv 1409.1556..
[49] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778..

Rashmi Kethireddy received Bachelor of Technology degree from Kakatiya Institute of Technology and Science, Warangal, India, in 2011, with a specialization in Information Technology (IT). She then worked in IT services for a period of two years. Post that, she received Master of Technology degree from Osmania University, Hyderabad, India, in 2017, with a specialization in Computer Science Engineering. She qualified for University Grant Commission National Eligibility Test (UGC-NET) and hence was awarded with Junior Research Fellowship (JRF) and Senior Research Fellowship (SRF). She is currently a Ph.D., scholar at International Institute of Information Technology, Hyderabad (IIIT-H). Her research interests include speech signal processing, acoustic analysis, machine learning, speech dialectal challenges, and speech dialect identification.

Sudarsana Reddy Kadiri is currently a Research Fellow with the Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland. Previously, he was a Postdoctoral Researcher at the same department in Aalto University from 2019–2021. He received the Ph.D. degree from the Department of electronics and communication engineering (ECE), International Institute of Information Technology, Hyderabad (IIIT-H), Hyderabad, in 2018, and B.Tech. degree from Jawaharlal Nehru Technological University (JNTU), Hyderabad, India, in 2011, with a specialization in ECE. He was a Teaching Assistant for several courses at IIIT-H from 2012 to 2018. Since 2019, he has been involved in teaching and mentoring activities at Aalto University, Espoo, Finland. His research interests include signal processing, speech analysis, speech synthesis, paralinguistics, affective computing, voice pathologies, machine learning, and auditory neuroscience. He has published over 50 research papers in peer-reviewed journals and conferences in these areas, and he is the reviewer for several reputed journals and conferences. Dr. Kadiri was awarded the Tata Consultancy Services (TCS) Fellowship for his Ph.D. degree.

Dr. Suryakanth V Gangashetty is a faculty member at KL University Green Field Vaddeswaram, Guntur District, Andhra Pradesh, India. He completed his PhD (in Neural Network Models for Recognition of Consonant–Vowel Units of Speech in Multiple Languages) from IIT Madras in 2005. Before joining KL University, he worked as a member of faculty at IIIT Hyderabad, Telangana, from 2006 to 2020. Previously he has worked as a Senior Project Officer at Speech and Vision Laboratory, IIT Madras. He has worked as a member of faculty at BIET Davangere Karnataka, from 1991 to 1999. He has also worked as a visiting research scholar at OGI Portland (USA) for three months during the summer of 2001. He has done his post-doctoral studies (PDF) at Carnegie Mellon University (CMU) Pittsburgh (PA, USA) during April 2007 to July 2008. He is an author of about 150 papers published in national as well as international journals, conferences, and edited volumes. He is a life member of the CSI, IE, IUPRAI, ASI, IETE, ORSI, and ISTE. He has reviewed papers for reputed journals and conferences. His research interests include Speech Processing, Neural Networks, Machine Learning, Natural Language Processing, Artificial Intelligence. He was the local Organizing Chair for the INTERSPEECH-2018 conference which happened in India in September 2018 held at Hyderabad.