Towards Relevance and Sequence Modeling in Language Recognition

Bharat Padi, Anand Mohan and Sriram Ganapathy, *Senior Member, IEEE*

***Abstract*—The task of automatic language identification (LID) involving multiple dialects of the same language family in the presence of noise is a challenging problem. In these scenarios, the identity of the language/dialect may be reliably present only in parts of the temporal sequence of the speech signal. The conventional approaches to LID (and for speaker recognition) ignore the sequence information by extracting long-term statistical summary of the recording assuming an independence of the feature frames. In this paper, we propose a neural network framework utilizing short-sequence information in language recognition. In particular, a new model is proposed for incorporating relevance in language recognition, where parts of speech data are weighted more based on their relevance for the language recognition task. This relevance weighting is achieved using the bidirectional long short-term memory (BLSTM) network with attention modeling. We explore two approaches, the first approach uses segment level i-vector/x-vector representations that are aggregated in the neural model and the second approach where the acoustic features are directly modeled in an end-to-end neural model. Experiments are performed using the language recognition task in NIST LRE 2017 Challenge using clean, noisy and multi-speaker speech data as well as in the RATS language recognition corpus. In these experiments on noisy LRE tasks as well as the RATS dataset, the proposed approach yields significant improvements over the conventional i-vector/x-vector based language recognition approaches as well as with other previous models incorporating sequence information.**

***Index Terms*—Language Recognition, i-vectors, Long-Short Term Memory (LSTM) networks, Sequence Modeling, Attention Networks.**

I. Introduction

THE problem of recognizing the spoken language of a given audio segment is of considerable interest for several commercial applications like speech translation [1], multi-lingual speech recognition [2], document retrieval [3] as well as in defense and surveillance applications [4]. In the recent years, several advances in speech signal modeling, most prominent of them being the application of factor analysis methods, have contributed to improving the performance of language recognition systems [5]. However, the task is still of considerable challenge when the recognition involves multiple dialects of the same language (like the recent NIST Language Recognition Evaluation (LRE) 2017 challenge). Further, the language identification (LID) performance is degraded in the presence of noise and other artifacts, such as in the robust automatic transcription of speech (RATS) databases [4], [6], [7]. In this paper, we propose a modeling framework to address the challenge of noise in language recognition.

Traditionally, phoneme recognition followed by language modeling (PRLM) was one of the popular methods for automatic LID task [8], [9]. In the recent past, the use of deep neural network (DNN) based posterior features were attempted for LID [10]. The bottleneck features based on the acoustic model of a speech recognition system had also shown promising results for noisy language recognition [11].

The development of i-vectors as one of the primary representations for LID was first introduced in [12]. The i-vectors are features of fixed dimension derived from variable length speech utterances using a background model [13]. The background model can be a Gaussian mixture model (GMM) [14] or a DNN model [11], [15]. The i-vectors extracted from the training data are used to train classifiers such as support vector machines (SVMs) which perform the task of language identification [16], [17].

One of the main drawbacks of the i-vector representations [12] and the recently proposed x-vector representations [18] is the long term summarization of the audio signal. Even in the extraction of x-vector embeddings using the TDNN models [19], the temporal context of 15 frames (about 150msec.) is alone used in the forward propagation of frame level features. For tasks like dialect identification in the presence of noise and other artifacts, some regions of audio may be more reliable than the rest. We hypothesize that there is a need to extract information only from the relevant regions of the audio signal rather than the long-term summary of the signal for the task of noisy LID. The attention approach in neural network modeling was originally proposed for neural machine translation [20] and image captioning [21]. The attention approach to speech recognition was initially investigated for phoneme recognition [22] where the issues in dealing with variable length speech sequences were identified. Recently, the attention based models have also been applied for end-to-end speech recognition tasks [23]–[25]. The end-to-end approaches to language recognition have been explored with long short term memory (LSTM) networks and with DNNs [26]. A recent approach using curriculum learning had also been applied for noise robust language recognition [27]. However, the state-of-the-art language recognition systems using large scale NIST language recognition evaluation (LRE) challenges, continue to use the i-vector/x-vector based approaches with support vector

This work was partly supported by grants from Department of Science and Technology, Early Career Award (ECR-01341/2017) and the Extra Mural Grant (EMR-2016/007934).

Bharat Padi is with the mind.ai, Bangalore. This work was performed when he was a student at the LEAP lab.

Anand Mohan is with Amazon Alexa Lab, Bangalore, India. This work was performed when he was a student at the LEAP lab.

Sriram Ganapathy is with the Learning and Extraction of Acoustic Patterns (LEAP) lab, Electrical Engineering, Indian Institute of Science, Bengaluru, India, 560012

machine classifier [28].

In this paper, we propose an attention based framework for language recognition to perform relevance weighting of the audio signal region in the presence of noise. The term relevance used in this work corresponds to the relative importance of short regions of audio (1000msec. chunks) for determining the language/dialect of the given utterance. For neural modeling of relevance, we explore two approaches,

i Modeling short segment statistics from i-vector representations using a deep bidirectional LSTM model [29]. We refer to this as the i/x−BLSTM model.

ii Modeling the audio features directly using end-to-end deep recurrent model with hierarchical gated recurrent units. We refer to this model as HGRU [30].

The rest of the paper is organized as follows. The state-of-the-art systems using i-vector based model and the LSTM based neural network approach to end-to-end language recognition are discussed in Sec. II-A and Sec. II-B respectively. Then, we discuss several approaches for incorporating relevance using short-term i-vector representations in Sec. III. In Sec. III-B, we discuss the proposed hybrid model of using short-term i-vector representations in a neural framework for language recognition. Sec. III-C details the proposed end-to-end framework for language recognition. The details of the experimental set and the various datasets used is given in Sec. IV. The LID results of various language recognition tasks on LRE 2017 and RATS datasets are reported in Sec. V. This is followed by an analysis of the robustness of the proposed approaches in the presence of noise in Sec.VI. In Sec. VII, we summarize the important contributions of this paper.

II. Relevant Prior Work

A. i-vector/x-vector SVM LID system

The i-vectors constitute the widely used features for language recognition [12]. A Gaussian Mixture Universal Background Model (GMM-UBM) is obtained by pooling the front end features from all the utterances in the train dataset. The means of the GMM are adapted to each utterance using the Baum-Welch (BW) statistics of the front-end features. A Total Variability Model (TVM) is assumed as a generative model for the adapted GMM mean supervector, which is given by,

$$\boldsymbol{M}(s) = \boldsymbol{M}_0 + T\boldsymbol{y}(s),$$

where $\boldsymbol{M}_0$ and $\boldsymbol{M}(s)$ are the UBM mean supervector and the adapted mean supervector of recording s respectively. Here, T is a rectangular matrix, and $\boldsymbol{y}(s) \sim \mathcal{N}(\boldsymbol{0}, I)$ is a latent variable. The *maximum aposteriori* (MAP) estimate of $\boldsymbol{y}(s)$ given the front-end features of the recording s is called the i-vector of recording $\boldsymbol{y}^*(s)$. The i-vectors extracted for each of the speech files are processed with length normalization [31] and dimensionality reduction is performed using linear discriminant analysis (LDA).

Recently, x-vector representations have been developed for speaker and language recognition [18], [19]. The x-vector model is based on time delay neural network (TDNN) where the initial layers operate on frame level features. The higher layers convert these features to segment level by summarizing mean and standard deviation of frame level features. This is followed by 2 fully connected hidden layers and the entire neural network model is trained to classify languages. The utterance level features before the last fully connected hidden layer are used as embeddings and they are termed as x-vectors. The x-vector features can be used instead of i-vector embeddings in a SVM classifier for language recognition [19].

B. Long Short Term Memory Neural Network LID system

Long Short-Term Memory networks (LSTM) [32], [33] are designed to address the issue of learning long-term dependencies in recurrent neural networks [34]. In [35], [36], the authors have proposed an Long Short Term Memory Recurrent Neural Network (LSTM-RNN) based end-to-end model for exploiting temporal information useful for LID. In [35] authors have shown that the LSTM model outperforms a baseline i-vector system and a deep neural network model on short duration (3s) test segments. In a recent work, the authors of [36] show that neural networks are useful even for challenging NIST datasets where the LSTM model outperforms i-vector baseline on short duration (3 sec.) test segments but performs relatively worse on longer duration test segments (10 sec., 30 sec.).

III. Relevance In Language Recognition

We propose three different approaches for incorporating relevance and sequence information in language recognition.

A. Relevance Weighted Baum-Welch statistics

The main motivation in this approach is the use of inverse entropy as measure of confidence/relevance. The entropy of class posterior distribution carriers information regarding the uncertainty of the classification. Given a discrete class posterior distribution, the higher the entropy the lower the confidence of the decision. Misra et. al [37] had shown that the entropy increases with decrease in SNR as the posterior probability distribution becomes more and more uniform when the noise level increases. If two segments $s1, s2$ of the audio contain two different SNR values and if the corresponding i-vector representations be denoted as y_1 and y_2, the inverse of the entropy of the corresponding language posteriors $H(l|y_1)$ and $H(l|y_2)$ provides a measure of the reliability of the decisions from each of audio regions. The optimal language decision using the information from the two audio regions can be taken by using inverse entropy weighted linear combination of the posteriors [37], [38],

$$l^* = argmax_l \big[w_1 p(l|y_1) + w_2 p(l|y_2)\big] \tag{1}$$

where, w_1 and w_2 are weights given by

$$w_1 = \frac{\frac{1}{H(l|y_1)}}{\frac{1}{H(l|y_1)} + \frac{1}{H(l|y_2)}} \tag{2}$$

and $w_2 = 1 - w_1$. In the proposed work, the relevance weighting of the Baum-Welch statistics is used to provide the optimal statistics for i-vector estimation (similar to SAD probability estimation done in [39]).

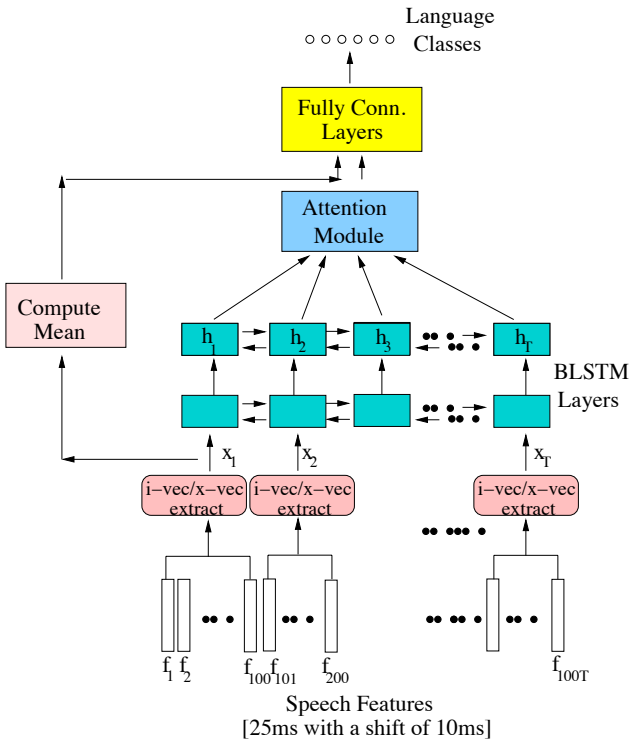


Fig. 1: Proposed BLSTM model using 1000 msec. i-vector/x-vector features for language recognition.

In our proposed approach, we use short term i-vectors that are labeled individually according to the label of the corresponding utterance. These 1000 msec. i-vectors from train and development datasets are then used to train a feed forward deep neural network (DNN). The DNN has an input layer of 500 dimensions with 3 hidden layers having 1024 dimensions and rectified linear unit (ReLU) non-linearity. The output layer was L dimensions where L is the number of language classes.

Once the DNN model is trained, for each non-overlapping short term i-vector $\boldsymbol{y}_i$ from every utterance, the information entropy H_i is computed from the posteriors from the DNN model where i is the index of 1000 msec. i-vector. We use the entropy measure H_i to compute a relevance parameter γ_i. The value of γ_i changes only at 1 sec. intervals. The value of γ_i is obtained from H_i as,

$$\gamma_i = \begin{cases} 1 & \text{when } H_i < H_{\mathsf{min}} \\ \frac{H_{\mathsf{max}} - H_i}{H_{\mathsf{max}} - H_{\mathsf{min}}} & \text{when } H_{\mathsf{min}} \leq H_i \leq H_{\mathsf{max}} \\ 0 & \text{when } H_i > H_{\mathsf{max}} \end{cases} \tag{3}$$

where H_{max} and H_{min} are hyper-parameters. The zeroth and first-order Baum-Welch statistics used for i-vector extraction are now modified as:

$$N_c(s) = \sum_{i=1}^{H(s)} \gamma_i \, \mathrm{p}(c \,|\, \boldsymbol{x}_i) \tag{4}$$

$$\boldsymbol{F}_{X,c}(s) = \sum_{i=1}^{H(s)} \gamma_i \, \mathrm{p}(c \mid \boldsymbol{x}_i)(\boldsymbol{x}_i - \boldsymbol{\mu}_c) \tag{5}$$

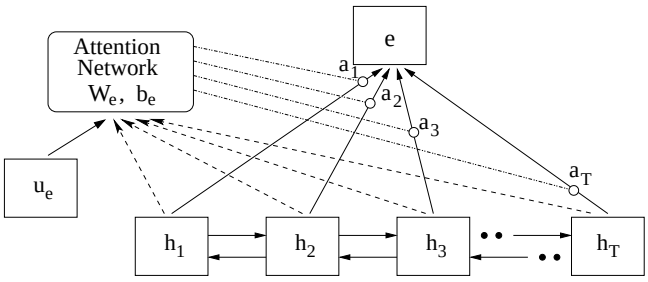


Fig. 2: Attention modeling in the proposed i/x-BLSTM/HGRU model.

B. Hybrid i-vector/x-vector BLSTM model

This approach is shown in Fig. 1. The short-term i-vectors/x-vectors, extracted every 200 msec. from overlapping windows of 1000 msec. duration (100 frames of acoustic features $f_1, f_2, ...$ extracted with 10 msec. shift) are modeled using a bidirectional LSTM model (BLSTM) with attention. The input to the BLSTM is the variable length sequence of short-term i-vectors/x-vectors. The LSTM architecture that we use in this paper contains two layers with 256 memory cells in each layer[1]. We propose to use an attention model to weight the 1 sec. representations based on their relevance to the language classification task. The attention method [20] provides an efficient way to aggregate the sequence of 1 sec. vectors. The attention mechanism used in this work is shown in Fig. 2. The model implements the following set of equations,

$$\mathbf{u}_t = tanh(\mathbf{W}_e \mathbf{h}_t + \ \mathbf{b}_e) \tag{6}$$

$$a_t = \frac{exp(\mathbf{u}_t^T \mathbf{u}_e)}{\sum_t exp(\mathbf{u}_t^T \mathbf{u}_e)} \tag{7}$$

$$\mathbf{e} = \sum_t a_t \mathbf{h}_t \tag{8}$$

Here, $\mathbf{W}_e$, $\mathbf{b}_e$ are the weights and the bias of the attention module which are learned in training process along with the vector $\mathbf{u}_e$. The output $\mathbf{e}$ denotes the fixed dimensional embedding from the input sequence. The attention module based on the similarity of $\mathbf{u}_t$ with $\mathbf{u}_e$ assigns normalized weights $\mathbf{a}_t$ using a softmax function. The utterance level representation $\mathbf{e}$ is mapped to the final language targets through a layer of fully connected network (FCN) with 512 dimensions and ReLU non-linearity. The output layer has a softmax non-linearity and the model is trained with cross entropy loss using stochastic gradient descent [40] and is implemented in TensorFlow [41].

C. End-to-End Hierarchical GRU model

We propose to use a simplified version of the LSTM function, the Gated Recurrent Unit (GRU) architecture [42] in the end-to-end model. The GRUs combine the input and forget gate into one single gate and in many tasks, the GRUs that have a relatively smaller number of parameters are shown to achieve or improve over the performance of LSTMs [43].

[1]All models proposed in this work are available at https://github.com/iiscleap/lre-relevance-weighting

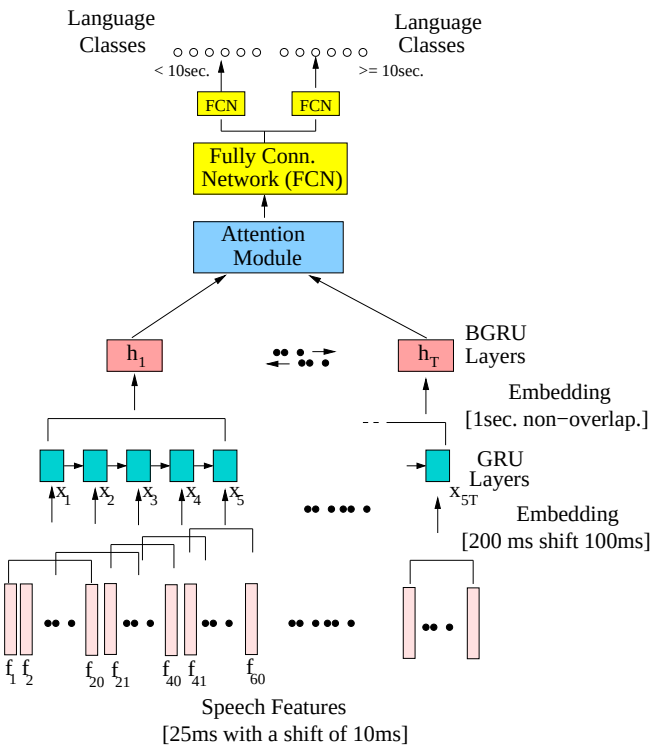


Fig. 3: End-to-end Hierarchical GRU RNN with attention module and duration dependent target layers.

Even with LSTM/GRU models, the direct modeling of long sequences can be cumbersome [36]. In order to model such long sequences, we propose a novel hierarchical bidirectional GRU [30] network with attention in this paper.

The block schematic of the proposed model is given in Fig. 3. At the first layer, a 256 cell unidirectional GRU block accumulates information across a window of 200 msec. i.e, a sequence of 20 feature vectors with a shift of 100 msec. over the entire input sequence. The output from the first layer is a sequence of vectors that are sampled every 100 msec. with each vector representing information from overlapping 200 msec. segments of input speech. This is then fed to the second layer of GRU block with 512 cells where the information is accumulated over a window of 1 sec. The accumulated 1 sec. vectors from second layer are fed to the final bidirectional GRU layer [44] with 512 cells in the third layer.

The output of the three layer hierarchical GRU model contains representations at 1 sec. level. The rest of model framework is similar to the $i\text{-}BLSTM$ model (Sec. III-B). For the end-to-end model, we found that the distribution of the embeddings from the attention network is quite different for short and long duration inputs. Hence, we propose to use two separate target layers, one for short duration inputs that are of the order of 3 sec. duration and other one with longer duration input sequences that are 10 sec. or above. The entire network is trained using Adam optimization and Back Propagation Through Time (BPTT) algorithm [40].

TABLE I: LRE17 training set : language clusters, target languages and total number of hours.

Cluster	Target Languages	Hours
Arabic	Egyptian Arabic (ara-arz)	190.9
	Iraqi Arabic (ara-acm)	130.8
	Levantine Arabic (ara-apc)	440.7
	Maghrebi Arabic (ara-ary)	81.8
Chinese	Mandarin (zho-cmn)	379.4
	Min Nan (zho-nan)	13.3
English	British English (eng-gbr)	4.8
	General American English (eng-usg)	327.7
Slavic	Polish (qsl-pol)	59.3
	Russian (qsl-rus)	69.5
Iberian	Caribbean Spanish (spa-car)	166.3
	European Spanish (spa-eur)	24.7
	Latin American Continental Spanish (spa-lac)	175.9
	Brazilian Portuguese (por-brz)	4.1

IV. Experimental set up

A. Feature Extraction

All the systems in this paper use the same front end features. The features are the Deep Neural Network (DNN) Bottleneck (BN) features [11], [28]. We extract BN features from a feed forward deep neural network trained for automatic speech recognition (ASR) using Kaldi [45] framework. The ASR-DNN was trained using 39 ($13 + \Delta + \Delta\Delta$) dimensional mel frequency cepstral coefficient (MFCC) features with 10 msec. frame shift and 25 msec. windows. The ASR-DNN was trained on speech from combined Switchboard (SWB1) and Fisher corpora (about 2000 hours of labeled audio). The ASR-DNN model uses 7 hidden layers with ReLU activation with layer-wise batch normalization. We use the last hidden layer output of size 80.

B. Datasets

1) NIST LRE 2017 Dataset: The details of the LRE17 dataset is given in Table I. The LRE17 training data (LDC2017E22) has five major language clusters with 14 target dialects with a total duration of 2069 hours in 16205 files. The development dataset consists of 3661 files which contain 253 hours of audio and the evaluation dataset consists of 25451 files with 1065 hours of audio. The development and evaluation datasets are further partitioned into utterances of durations 3 sec, 10 sec, 30 sec. or 1000 sec. The datasets contain conversational telephone speech (CTS) and broadcast narrow band speech (BNBS) and speech extracted from videos or video speech (VS) (the 1000 sec. files). We have not used the development data in training the embedding models (i-vector/x-vector) or the back-end models.

In addition to the standard LRE17 test set, we test the robustness of the various LID systems using noisy versions of the evaluation data for the 10 sec. recording conditions. We use four different noise types (Babble, Subway, Airport and Street) from the Aurora-4 corpus and these are added to the audio signal at different signal-to-noise ratio (SNR) like 5, 10, 15 and 20 dB using the filter and noise adding tool (FANT) [46]. While adding the noise, either the entire audio file is corrupted with noise ($Noisy$ cond.) or only the first half of the audio file is corrupted with noise ($Partial\ Noise.$). The

TABLE II: Performance of reference and the developed systems on the NIST LRE17 evaluation dataset in terms of percentage accuracy, C_{avg} and EER. The performance of the x-vector systems are shown in parentheses. The x-vector systems were evaluated only for 3,10 and 30s conditions. The x-post system uses the x-vector network directly without the proposed BLSTM backend model.

Dur.	i-LDA-SVM (x-LDA-SVM)	LSTM [36]	DNN-Attn [27]	RWBW	i-BLSTM (x-BLSTM)	HGRU	x-post.	x-BLSTM-E2E
Accuracy (%)								
3	53.84 (46.14)	54.7	54.6	51.37	54.80 (56.99)	55.13	(60.55)	(**60.2**)
10	72.36 (73.19)	72.1	72.5	68.42	**75.89** (71.60)	74.06	(69.82)	(70.5)
30	82.98 (**85.09**)	76.1	79.7	77.59	82.27 (76.02)	82.98	(72.02)	(74.2)
1000	56.23	42.8	50.2	**58.48**	54.07	53.53	-	-
overall	67.86	64.3	66.36	64.78	**68.65**	68.48	-	-
C_{avg}								
3	0.53 (0.49)	0.55	0.53	0.58	0.50 (0.46)	0.55	(**0.45**)	(0.46)
10	0.27 (**0.21**)	0.35	0.28	0.35	0.26 (0.28)	0.32	(0.32)	(0.32)
30	0.13 (**0.10**)	0.28	0.20	0.21	0.18 (0.23)	0.23	(0.28)	(0.28)
1000	0.54	0.79	0.61	0.51	**0.50**	0.62	-	-
overall	0.37	0.48	0.40	0.40	**0.36**	0.42	-	-
EER (%)								
3	**13.40** (16.92)	15.39	14.98	15.60	15.47 (13.50)	15.33	(13.70)	(13.45)
10	6.47 (**5.98**)	8.70	7.05	7.84	6.32 (7.33)	7.49	(9.20)	(8.53)
30	3.50 (**2.75**)	7.25	4.73	4.61	3.67 (6.28)	4.93	(8.11)	(7.12)
1000	15.35	26.27	15.72	**14.20**	14.71	17.02	-	-
overall	**9.26**	14.38	10.80	10.34	9.65	10.79	-	-

latter condition is considered to simulate non-stationary test conditions where the noise characteristics can change within the course of recording duration.

2) RATS LID Dataset: The DARPA Robust Automatic Transcription of Speech (RATS) [4] program targets the development of speech systems operating on highly distorted speech recorded over "degraded" radio channels. The data used here consists of recordings obtained from re-transmitting a clean signal over eight different radio channel types, where each channel introduces a unique degradation mode specific to the device and modulation characteristics [4]. For the language identification (LID) task, the performance is degraded due to the short segment duration of the speech recordings in addition to the significant amount of channel noise [47].

The training data for the RATS experiments consist of 20000 recordings (about 1600 hours of audio) from five target languages (Arabic, Pashto, Dari, Farsi and Urdu) as well as from several other non-target languages. We have used 6 out of 8 given channels (channels B-G) for training and testing purposes. The development and the evaluation data consists of 5663 and 14757 recordings respectively from the above 6 channels. We also evaluate the models on sampled 3 sec, 10 sec. and 30 sec. chunks of voiced data from the full length evaluation files.

C. Evaluation Metrics

The performance is measured using the primary metric described in the evaluation plan of NIST LRE 2017 C_{avg} [28], Equal Error Rate (EER) and accuracy.

- $\mathbf{C_{avg}}$ - The pair-wise language recognition performance will be computed for all the target-language/non-target-language pairs (L_T, L_N). An average performance cost for each system is computed as

$$C_{avg}(\beta) = \frac{1}{N_L} \sum_{L_T} \left\{ P_m(L_T) + \sum_{L_N} \frac{\beta P_f(L_T, L_N)}{N_L - 1} \right\}$$

  where P_m is the probability of miss detection, P_f is the probability of false alarm, which are computed by applying detection threshold of $\log \beta$ to the system scores. The primary metric used for LRE17 is the average cost of two C_{avg} measures obtained using $\beta_1 = 1$, $\beta_2 = 9$.
- **EER** - Equal Error Rate is measured for individual languages and then their average is computed. Note that, the EER and C_{avg} consider the LID system as detection problem.
- **Accuracy** - The accuracy is measure in a classification setting. It considers the LID systems as multi-class learning problem (14 classes in LRE17 dataset and 6 classes in the RATS evaluation).

D. i-vector/x-vector baseline system

Once the BN features are extracted from the ASR-DNN for the LID data, a speech activity detection (SAD) algorithm was applied to remove the unvoiced frames [48]. We use the implementation of SAD from the Voicebox toolkit [49]. This was followed by cepstral mean variance normalization (CMVN) done over each utterance, followed by a sliding window CMVN over a sliding window of 3 sec. The number of UBM mixture components was set to $C = 2048$ and the dimension of the total variability space was fixed to be $R = 500$. The i-vectors are processed with within class covariance normalization (WCCN) technique [50] and length normalized. The dimension of the i-vectors is then reduced to 13 and 5 for LRE17 and RATS datasets respectively using linear discriminant analysis (LDA) and modeled using support vector machines (LDA-SVM). The i-vector SVM system implementation follows the recipe provided by the NIST LRE system [28].

The x-vector baseline system is implemented using the TDNN model architecture described in [19]. The x-vector system consists of 5 layers of time delay neural network

TABLE III: Performance of the systems when evaluated on data (10 sec. recording condition) corrupted with noise at various SNR levels. The performance of the x-vector systems are shown in parentheses.

SNR	i-LDA-SVM (x-LDA-SVM)	LSTM [36]	DNN-Attn [27]	i-BLSTM (x-BLSTM)	HGRU	(x-post)	(x-BLSTM-E2E)
Accuracy (%)							
5dB	47.28 (49.96)	48.36	46.35	51.58 (55.44)	45.78	(**58.13**)	(57.41)
10dB	53.42 (58.43)	56.30	54.86	59.86 (62.65)	53.92	(63.57)	(**63.73**)
15dB	57.47 (63.20)	61.63	61.58	64.30 (66.27)	60.20	(65.88)	(**66.94**)
20dB	59.64 (66.66)	65.28	65.76	67.72 (68.24)	64.21	(67.40)	(**68.36**)
Avg.	54.45 (59.56)	57.89	57.13	60.87 (63.15)	56.02	(63.74)	(**64.36**)
Cavg							
5dB	0.62 (0.71)	0.65	0.60	0.59 (**0.47**)	0.68	(0.47)	(0.50)
10dB	0.53 (0.63)	0.54	0.49	0.48 (**0.38**)	0.58	(0.40)	(0.41)
15dB	0.48 (0.58)	0.47	0.41	0.42 (**0.33**)	0.48	(0.36)	(0.37)
20dB	0.45 (0.54)	0.42	0.37	0.37 (**0.31**)	0.43	(0.34)	(0.35)
Avg.	0.52 (0.62)	0.52	0.47	0.47 (**0.37**)	0.54	(0.39)	(0.41)
EER (%)							
5dB	22.78 (26.52)	19.79	17.95	17.01 (**14.44**)	19.16	(16.25)	(15.54)
10dB	18.20 (21.97)	15.88	13.27	12.90 (**11.11**)	15.61	(12.64)	(11.94)
15dB	15.63 (19.30)	12.87	10.88	11.05 (**9.26**)	13.19	(10.95)	(10.07)
20dB	13.99 (17.87)	11.11	9.43	9.44 (**8.37**)	11.69	(10.11)	(9.20)
Avg.	17.65 (21.41)	14.91	12.88	12.60 (**10.80**)	14.91	(12.49)	(11.69)

(TDNN) architecture followed by statistics (mean and std. dev.) pooling layer which converts frame level features to utterance level features. These segment level statistics are fed through two layers of feed forward network with 512 hidden dimensions. The x-vectors are the embeddings from the affine layer after the statistics pooling (output of layer 6 before the non-linearity with 512 dimensions). The target for the x-vector model are the language labels. We use the LRE2017 training data along with data augmentation to train the x-vector model. The model used a five fold data augmentation procedure with room reverberation as well as the additive noises like music, noise and babble (MUSAN corpus [51]) at SNR values of $5, 10, 15, 20$ dB. The entire model is trained using the Kaldi recipe [45]. We report the results using x-vector features for the LRE2017 evaluation in Table II. We also report the performance of the x-vector system directly using the posteriors from the trained network (denoted as x-post).

E. DNN and LSTM baseline

We implement the best performing LSTM model [36], which is a two layer LSTM with 512 units in each layer followed by an output softmax layer as a baseline end to end system. We also compare the performance of proposed models with a DNN with attention developed recently [27].

F. Proposed Relevance Models for LID

The i-BLSTM model uses the i-vector extraction setup similar to the one used in the baseline system. The audio recordings are chunked into 1000 msec. which are shifted every 200 msec. The segment i-vectors are used to train the BLSTM model on the training data. We use $80,000$ recordings of 15 sec. duration from the LRE training dataset and the model is trained using a cross entropy loss. For training the proposed end-to-end HGRU model on the LRE17 data, audio snippets of duration ranging from 3 sec. to 30 sec. are randomly sampled from the training data. A similar kind of setup is used for RATS dataset, where around $100,000$ samples of 15 sec. duration is sampled for training the i-BLSTM model and random sampling of 3 sec. and 30 sec. is done for the end-to-end HGRU model. Finally, we also experiment with replacing the i-vector features with the x-vector features for the BLSTM model (this model is refered to as x-BLSTM system).

V. Language Recognition Results

A. LRE Evaluation Results

The results for LRE evaluations are listed in Table II. The relevance weighted models proposed in this paper are the relevance weighted Baum-Welch statistics (RWBW), the hybrid i-vector BLSTM model (i-BLSTM) and the hierarchical GRU model (HGRU).

As seen in Table II, in terms of the C_{avg}, the baseline neural models like LSTM [36] and DNN with attention [27] perform comparatively well on short durations (3 sec. and 10 sec.). However, on longer duration of 30 sec. and 1000 sec, the i-vector based LDA-SVM system provides the best performance. This indicates that the limitations of the previous models to incorporate the long term dependencies. The proposed relevance based approaches compare favorably with the i-vector SVM baseline in all conditions. The RWBW system provides the best performance on the long duration (1000 sec.) condition as there are a large number of short segment i-vectors in these recordings. However, in short duration conditions, the RWBW is inferior to other proposed methods. The HGRU model performs similar to the DNN-attn model proposed previously. The best system in terms of the C_{avg}. and the EER metric is the x-BLSTM system.

B. Results on Noisy LRE Data

The results with various noisy versions of the evaluation data are shown in Table III and Table IV. For the experiments with noisy data (Table III), among the various baseline systems, the LSTM model [36] gives the best robustness. The proposed approach of hybrid i-BLSTM model provides significant improvements over all models considered for all

TABLE IV: Performance of the systems with partial noise (50% of the utterance, 10 sec. recording condition) at various SNR levels. The x-vector based systems are shown in parentheses.

SNR	i-LDA-SVM (x-LDA-SVM)	LSTM [36]	DNN-Attn [27]	i-BLSTM (x-BLSTM)	HGRU	(x-post)	(x-BLSTM-E2E)
Accuracy (%)							
5dB	53.16 (57.36)	56.50	54.64	59.79 (61.41)	55.32	(58.85)	(**61.68**)
10dB	55.57 (61.07)	60.42	58.22	63.02 (63.56)	60.20	(61.44)	(**63.64**)
15dB	58.33 (64.10)	62.61	61.61	**65.90** (65.66)	63.16	(63.05)	(65.60)
20dB	59.56 (65.53)	64.61	63.55	**68.16** (66.51)	65.65	(63.79)	(66.16)
Avg.	56.65 (62.01)	61.04	59.50	64.20 (**64.29**)	61.08	(61.78)	(64.27)
Cavg							
5dB	0.54 (0.66)	0.52	0.50	0.47 (**0.38**)	0.57	(0.45)	(0.44)
10dB	0.50 (0.61)	0.47	0.44	0.43 (**0.35**)	0.51	(0.41)	(0.41)
15dB	0.47 (0.58)	0.44	0.41	0.39 (**0.33**)	0.47	(0.38)	(0.38)
20dB	0.45 (0.56)	0.42	0.39	0.37 (**0.32**)	0.44	(0.37)	(0.37)
Avg.	0.49	0.46	0.44	0.42 (**0.35**)	0.50	(0.40)	(0.40)
EER (%)							
5dB	17.56 (22.32)	14.59	13.28	12.30 (**10.67**)	15.36	(13.81)	(12.54)
10dB	15.79 (20.66)	13.10	11.52	10.84 (**9.75**)	13.05	(12.41)	(11.22)
15dB	14.32 (18.75)	11.80	10.45	9.60 (**9.04**)	11.89	(11.60)	(10.62)
20dB	13.70 (17.97)	11.20	9.64	8.95 (**8.59**)	11.08	(11.34)	(10.22)
Avg.	15.34 (19.92)	12.67	11.22	10.42 (**9.51**)	12.85	(12.29)	(11.15)

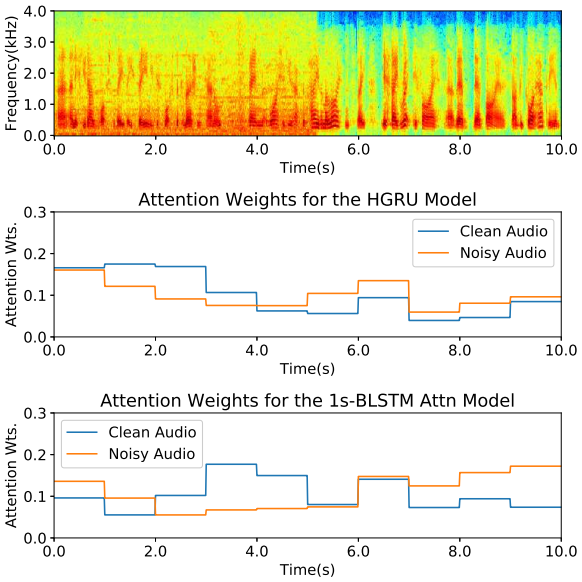


Fig. 4: Plot of a partially noisy (noise present in the first half of the audio) LRE-17 audio file spectrogram and the corresponding attention weights estimated every 1000 msec. for the proposed HGRU model and the i-BLSTM model.

SNR conditions. The hierarchical GRU model also improves over the LDA-SVM baseline but the performance of this system is inferior to the hybrid x-BLSTM system as well as the DNN-attn model [27]. On the average, the proposed x-BLSTM model shows a relative improvement of 28.8 % over the baseline i-LDA-SVM in terms of the accuracy measure.

On the partially noisy conditions (Table IV), the trends are similar. The baseline performance of the LDA-SVM system is improved by the DNN-attn model. However, the proposed i-BLSTM model improves over the previous neural network model based approaches for LID. On the average, the x-BLSTM model improves the baseline LDA-SVM relatively by 28.5% in term of the accuracy measure.

The reason for the counterintuitive lack of improvement for the partial noise over the full noise case for the baseline systems (comparing Table III and Table IV for x-post system results) may be attributed to the lack of ability of the baseline methods like x-post in modeling the non-stationarity of the signal (partial noise). The results for the proposed approach (x-BLSTM) in terms of [Acc]{C_avg}(EER) reported in Table III are [63.15]{0.37}(10.80). These results are improved for the partial noise case in Table IV as [64.29]{0.35}(9.51). We attribute this improvement to the ability of the attention modeling to re-weight the relevance to regions of the signal that are less noisy.

TABLE V: Performance of the baseline system and the proposed models on the RATS dataset in terms of accuracy, C_{avg} and EER.

Duration (sec.)	i-LDA-SVM	i-BLSTM	HGRU
Accuracy (%)			
3	65.39	64.16	**69.78**
10	77.46	79.07	**81.76**
30	85.38	87.12	**87.89**
120	92.22	**92.69**	91.43
Cavg			
3	1.12	1.04	**0.84**
10	0.81	0.63	**0.57**
30	0.57	**0.43**	**0.43**
120	0.40	**0.32**	0.35
EER (%)			
3	25.90	24.32	**19.56**
10	17.81	13.69	**12.23**
30	11.88	**8.61**	8.87
120	7.64	**5.74**	6.81

C. Results on RATS Dataset

Here, we compare the performance of the i-vector LDA-SVM with the proposed approaches of hybrid i-BLSTM system and the end to end HGRU model. As seen in the results, the proposed approaches yields significant improvements over the baseline system in terms of all the performance metrics and for all the duration conditions. The HGRU model is more efficient in the short duration conditions while the i-BLSTM

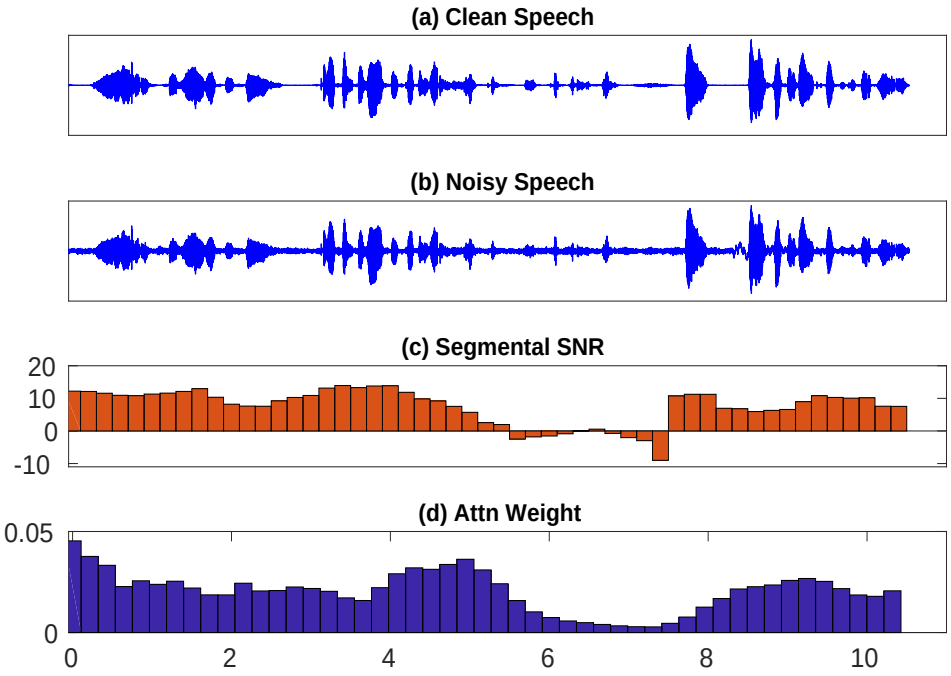


Fig. 5: Plot of (a) a clean LRE utterance from validation data, (b) the same utterance with 10 dB babble noise, (c) segment level SNR measured with 1000msec. windows shifted by 200msec. and (d) the corresponding attention weights estimated from the i-BLSTM model for 1000 msec. windows shifted by 200 msec.

model improves over the HGRU model for longer duration, full length (120 sec.) recording conditions.

D. Results with x-vector embeddings

From Table II, we observe that the x-vector based LDA-SVM improves over the baseline i-vector based system with similar back-end in terms of C_{avg} measure for all durations. The proposed BLSTM based backend improves the x-LDA-SVM system significantly for short duration (3sec. condition). However, for the longer durations, the baseline xvec. LDA-SVM provides better performance than the proposed model. The x-vector network posteriors directly used for LRE is also worse compared to the use of the SVM backend (especially for long durations).

The results on noisy LRE17 dataset and partially noisy dataset are given in Table III and Table IV respectively. As seen here, the x-vector embeddings provide noticeable improvements over the i-vector counterparts. Further, the proposed backend approach of x-BLSTM provides significant improvements over the baseline system both in the noisy and partially noisy case. The relative improvements in terms of C_{avg} metric over the baseline system ranges from $34-43$ % for noisy conditions and about 42 % for partially noisy conditions. It is also noteworthy that the proposed approach suffers from a performance degradation of less than 20% relative (in terms of C_{avg}) compared with clean conditions when then SNR is above 15 dB while the baseline x-LDA-SVM system suffers from more than 40 % degradation in relative performance. These results are also in alignment with i-vector based results.

VI. Discussion and Analysis

A. Importance of Relevance Modeling

The plot shown in Fig. 4 illustrates the spectrogram of a partially noisy (noise present in the first half of the audio recording) and the corresponding attention weights from the HGRU and the i-BLSTM model. The attention weights for the same audio recordings without any noise (clean) condition are also shown for reference. As seen in this plot, the attention weights are considerably changed due to the presence of the noise in the first half of the file. The model is able to be adaptive and focuses the language detection on the more reliable regions of the audio. The increased senstivity of the i-BLSTM system also potentially explains the improved performance of the i-BLSTM model in the experiments reported in Table III and Table IV.

Specifically, in the presence of noise, different parts of the signal have different signal-to-noise ratio values (as the speech signal is time-varying even if the noise is stationary). This can be visualized using the example given in Fig. 5. While the SNR for this entire speech utterance is 10 dB, the SNR measured at short-term segment level is highly time varying (ranging between 20 dB and -10 dB). The goal of the relevance models for LID proposed in this work is to deemphasize the contributions from the noisy regions while enhancing the contribution of the high SNR regions in making the decision about the language identity. An example of the

attention weights from the i-BLSTM model is also shown in Fig. 5 (d). As seen in this plot, the attention weights tend to track the short-term SNR thereby weighting the embeddings from the regions of high SNR with higher weights while reducing the weights associated with regions of low SNR.

B. End-to-end LID with x-vector

With use of the x-vector based embeddings in the proposed approach, we explore whether the proposed attention based BLSTM and the embedding x-vector model can be jointly trained. This model is referred to as the x-BLSTM E2E. The results for clean LRE17 test set are also reported in Table II and the noisy LRE results are reported in Table III and Table IV. As seen here, the end-to-end system improves over the x-BLSTM system for short durations while the performance is degraded for longer duration conditions. It is also important to note that the performance of the x-BLSTM-E2E is better than the x-post system which uses the x-vector network outputs directly for LRE. These results indicate that the attention based backend modeling is important for noisy LRE task.

VII. Conclusion

In this paper, we have proposed a new hybrid i-vector/x-vector BLSTM attention model (i/x-BLSTM) for language recognition where the sequence of 1000 msec. i-vectors/x-vectors are modeled in a bidirectional attention based network. A novel processing pipeline with hierarchical gated recurrent unit (HGRU) or x-vector BLSTM model is also proposed for end-to-end spoken language recognition. While, the conventional backend systems based on LDA-SVM classifiers ignore sequential speech information, the attention mechanism in the proposed model plays the role of relevance weighting, where portions of the speech signal more relevant to classification decision are given higher weightage. This relevance modeling is shown in particular to improve the robustness of the model for language recognition in noisy conditions. With several experiments on the noisy LRE dataset as well as the RATS dataset, we show the effectiveness of the proposed model. We also present a detailed analysis, which highlights the role of attention modeling in language recognition.

Acknowledgment

The authors would like to acknowledge Shreyas Ramoji, Satish Kumar and Vaishnavi Y for their help in setting up the LRE 2017 baseline, Omid Sadjadi for code fragments implementing the NIST LRE baseline system. The work reported in this paper was partly funded by grants from the Department of Science and Technology (EMR/2016/007934) and by Department of Atomic Energy (DAE/34/20/12/2018-BRNS/34088).

References

[1] Alex Waibel, Petra Geutner, L Mayfield Tomokiyo, Tanja Schultz, and Monika Woszczyna, "Multilinguality in speech and spoken language systems," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1297–1313, 2000.

[2] Tanja Schultz and Alex Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, no. 1-2, pp. 31–51, 2001.

[3] Ciprian Chelba, Timothy J Hazen, and Murat Saraclar, "Retrieval and browsing of spoken content," *IEEE Signal Processing Magazine*, vol. 25, no. 3, 2008.

[4] Kevin Walker and Stephanie Strassel, "The rats radio traffic collection system," in *Odyssey 2012-The Speaker and Language Recognition Workshop*, 2012.

[5] Haizhou Li, Bin Ma, and Kong Aik Lee, "Spoken language recognition: from fundamentals to practice," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1136–1159, 2013.

[6] Sriram Ganapathy, Mohamed Omar, and Jason Pelecanos, "Unsupervised channel adaptation for language identification using co-training," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 6857–6861.

[7] Mitchell McLaren, Diego Castan, and Luciana Ferrer, "Analyzing the effect of channel mismatch on the sri language recognition evaluation 2015 system," in *Proc. Odyssey: Speaker Lang. Recognit. Workshop*, 2016, pp. 188–195.

[8] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, pp. 31, 1996.

[9] Jiri Navratil, "Spoken language recognition-a step toward multilinguality in speech processing," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 6, pp. 678–685, 2001.

[10] Mhamed Faouzi BenZeghiba, Jean-Luc Gauvain, and Lori Lamel, "Phonotactic language recognition using MLP features.," in *Interspeech*, 2012.

[11] Fred Richardson, Douglas Reynolds, and Najim Dehak, "Deep neural network approaches to speaker and language recognition," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1671–1675, 2015.

[12] Najim Dehak, Pedro A Torres-Carrasquillo, Douglas Reynolds, and Reda Dehak, "Language recognition via i-vectors and dimensionality reduction," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[13] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[14] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[15] Yun Lei, Nicolas Scheffer, Luciana Ferrer, and Mitchell McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1695–1699.

[16] Sriram Ganapathy, Kyu Han, Samuel Thomas, Mohamed Omar, Maarten Van Segbroeck, and Shrikanth S Narayanan, "Robust language identification using convolutional neural network features," in *Fifteenth annual conference of the international speech communication association*, 2014.

[17] Bharat Padi, Shreyas Ramoji, Vaishnavi Yeruva, Satish Kumar, and Sriram Ganapathy, "The LEAP language recognition system for lre 2017 challenge-improvements and error analysis," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 31–38.

[18] David Snyder, Daniel Garcia-Romero, and Daniel Povey, "Time delay deep neural network-based universal background models for speaker recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 92–97.

[19] David Snyder, Daniel Garcia-Romero, Alan McCree, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "Spoken language recognition using x-vectors.," in *Odyssey*, 2018, pp. 105–111.

[20] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[21] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[22] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[23] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio, "End-to-end attention-based large vocabulary speech recognition," in *Acoustics, Speech and Signal Processing*

(ICASSP), 2016 IEEE International Conference on. IEEE, 2016, pp. 4945–4949.

[24] Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[25] Shubham Bansal, Karan Malhotra, and Sriram Ganapathy, "Speaker and language aware training for end-to-end asr," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 494–501.

[26] KV Mounika, Sivanand Achanta, HR Lakshmi, Suryakanth V Gangashetty, and Anil Kumar Vuppala, "An investigation of deep neural network architectures for language recognition in indian languages.," in *INTERSPEECH*, 2016, pp. 2930–2933.

[27] Ravi Kumar Vuddagiri, Hari Krishna Vydana, and Anil Kumar Vuppala, "Curriculum learning based approach for noise robust language identification using dnn with attention," *Expert Systems with Applications*, vol. 110, pp. 290–297, 2018.

[28] Seyed Omid Sadjadi et al., "The 2017 NIST language recognition evaluation," in *Proc. Odyssey*, Les Sables dÓlonne, France, June 2018.

[29] Bharat Padi, Anand Mohan, and Sriram Ganapathy, "Attention based hybrid i-vector blstm model for language recognition," *Proc. Interspeech 2019*, pp. 1263–1267, 2019.

[30] Bharat Padi, Anand Mohan, and Sriram Ganapathy, "End-to-end language recognition using attention based hierarchical gated recurrent unit models," in *Proc. ICASSP*, 2019.

[31] Daniel Garcia-Romero and Carol Y Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[32] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[33] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins, "Learning to forget: Continual prediction with LSTM," *IET*, 1999.

[34] Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, Jürgen Schmidhuber, et al., "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies," 2001.

[35] Ignacio Lopez-Moreno, Javier Gonzalez-Dominguez, Oldrich Plchot, David Martinez, Joaquin Gonzalez-Rodriguez, and Pedro Moreno, "Automatic language identification using deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 5337–5341.

[36] Ruben Zazo, Alicia Lozano-Diez, and Joaquin Gonzalez-Rodriguez, "Evaluation of an lstm-rnn system in different nist language recognition frameworks," in *Proc. of Odyssey 2016 Speaker and Language Recognition Workshop*. ATVS-UAM, June 2016.

[37] Hemant Misra, Hervé Bourlard, and Vivek Tyagi, "New entropy based combination rules in hmm/ann multi-stream asr," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).* IEEE, 2003, vol. 2, pp. II–741.

[38] Fabio Valente, "A novel criterion for classifiers combination in multi-stream speech recognition," *IEEE Signal Processing Letters*, vol. 16, no. 7, pp. 561–564, 2009.

[39] Mitchell McLaren, Martin Graciarena, and Yun Lei, "Softsad: Integrated frame-based speech confidence for speaker recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4694–4698.

[40] Paul J Werbos et al., "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[41] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[42] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[43] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.

[44] Mike Schuster, Kuldip K. Paliwal, and A. General, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, 1997.

[45] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, number EPFL-CONF-192584.

[46] H Guenter Hirsch, "Fant-filtering and noise adding tool," *Niederrhein University of Applied Sciences, http://dnt.-kr. hsnr. de/download. html*, 2005.

[47] K. J. Han, S. Ganapathy, M. Li, M. Omar, and S. Narayanan, "TRAP Language identification system for RATS phase II evaluation," in *Interspeech*. ISCA, 2013.

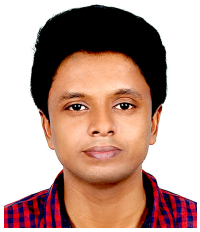
[48] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, pp. 1–3, 1999.

[49] N. Brummer, S. Cumani, O. Glembek, M. Karafiat, and P. Matejka, "Description and analysis of the Brno system for LRE2011," *Odyssey 2012-The Speaker and Language Recognition Workshop*, 2012.

[50] Andrew O Hatch, Sachin Kajarekar, and Andreas Stolcke, "Within-class covariance normalization for svm-based speaker recognition," in *Ninth international conference on spoken language processing*, 2006.

[51] David Snyder, Guoguo Chen, and Daniel Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

Bharat Padi received his Bachelor of Engineering in Electrical and Electronics Engineering from Andhra University, Visakhapatnam, India in 2009 and his Master's Degree in System Science and Automation from Indian Institute of Science (IISc), Bengaluru, India in 2018. Between 2009-2016, he worked as a systems engineer in Infosys Technologies and later as an Assistant Manager (Electrical) in Visakhapatnam steel plant. He is currently working in minds.ai as a neural network research engineer aiming at developing controllers for vehicles using deep reinforcement learning. His current research interests include developing deep end-end models for spoken language recognition and reinforcement learning. He is a member of the IEEE.



Anand Mohan is currently working as Applied Scientist in Alexa Speech team in Amazon, Bangalore, India. He received his Master's Degree in Artificial Intelligence from Indian Institute of Science (IISc), Bangalore, India in 2019 and his Bachelor's Degree in Electronics and Communication Engineering from National Institute of Technology, Calicut, India in 2015. His research interests include end-to-end ASR technologies, speaker and language identification.

Sriram Ganapathy is a faculty member at the Electrical Engineering, Indian Institute of Science, Bangalore, where he heads the activities of the learning and extraction of acoustic patterns (LEAP) lab. Prior to joining the Indian Institute of Science in 2016, he was a research staff member at the IBM Watson Research Center, Yorktown Heights. He received his Doctor of Philosophy from the Center for Language and Speech Processing, Johns Hopkins University in 2012. He obtained his Bachelor of Technology from College of Engineering, Trivandrum, India in 2004 and Master of Engineering from the Indian Institute of Science, Bangalore in 2006. He has also worked as a Research Assistant in Idiap Research Institute, Switzerland from 2006 to 2008.

At the LEAP lab, his research interests include signal processing, machine learning and deep learning methodologies for speech/speaker recognition and auditory neuroscience. He is a subject editor for the Speech Communications journal, member of ISCA and a senior member of the IEEE.