Attention based Hybrid I-vector BLSTM Model for Language Recognition

Bharat Padi[1*], *Anand Mohan*[2*], *Sriram Ganapathy*[2]

[1]minds.ai, Bengaluru, India.
[2]Learning and Extraction of Acoustic Patterns (LEAP) Lab, Electrical Engineering,
Indian Institute of Science, Bengaluru, India.

bharat@minds.ai, anandmohan@iisc.ac.in, sriramg@iisc.ac.in

Abstract

In this paper, a hybrid i-vector neural network framework (i-BLSTM) which models the sequence information present in a series of short segment i-vectors for the task of spoken language recognition (LRE) is proposed. A sequence of short segment i-vectors are extracted for every speech utterance and are then modeled using a bidirectional long short-term memory (BLSTM) recurrent neural network (RNN). Attention mechanism inside the neural network relevantly weights segments of the speech utterance and the model learns to give higher weights to parts of speech data which are more helpful to the classification task. The proposed framework performs better in short duration and noisy environments when compared with the conventional i-vector system. Experiments are performed on clean, noisy and multi-speaker speech data from NIST LRE 2017 and RATS language recognition corpus. In these experiments, the proposed approach yields significant improvements (relative improvements of 7.6 - 13% in terms of accuracy for noisy conditions) over the conventional i-vector based language recognition approach and also over an end-to-end LSTM-RNN based approach.

Index Terms: Spoken language recognition, short segment i-vectors, LSTM, attention

1. Introduction

The task of spoken language recognition has gained considerable interest in many applications in the recent years [1, 2, 3]. Even with the advancements in speech signal modeling methods like the factor analysis [4] which gave a significant boost to the performance of language identification (LID) systems, the primary challenge is with the utterances of short duration and when the task involves recognizing multiple dialects of same language family. The performance of the system is further degraded in the presence of noise and other artifacts as shown in the robust automatic transcription of speech (RATS) databases [3, 5, 6]. In this paper, we propose a short segment sequence modeling framework to address some of these challenges in LID system development.

The fixed dimensional embeddings known as i-vectors for a variable length speech utterances using a background model [7] for LID was introduced in [8]. The i-vectors are extracted from utterance level adaptations of a background model which can be a Gaussian mixture model (GMM) [9] or a DNN model [10]. They capture long term information contained in the speech utterance like the speaker and language. Once extracted, the i-vectors from the training data are used to train classifiers like the support vector machines (SVMs) to perform the task of language identification [11, 12].

One of the main drawbacks of the i-vector representations [8] and the recently proposed x-vector representations [10] is the global summarization of the audio signal. For tasks like dialect identification from short duration audio snippets, the information discerning the dialect/language may lie only in a small region of the audio signal (few words in the whole utterance). Also, in the presence of noise and other artifacts, some regions of audio may be more reliable than the rest. In these scenarios, the global summarization of the signal may suppress the key information in short audio snippets. We hypothesize that there is a need to model the relevant regions of the audio signal rather than the long-term summary of the signal for the task of LID.

Attention based approach in neural network modeling proposed in [13] for machine translation and [14] for image captioning has shown to provide relative importance to different temporal regions of the input sequence for sequence-to-sequence mapping tasks. Attention modeling approaches for speech recognition with variable length speech utterances were initially investigated for phoneme recognition in [15]. Recently, the attention based models have also been applied for end-to-end speech recognition [16, 17] and language recognition tasks [18]. For emotion recognition from speech, attention based models have been explored in [19, 20]. However, the state-of-art language recognition systems using large scale NIST language recognition evaluation (LRE) challenges, continue to use the i-vector based approaches with support vector machine classifier [21]. We propose a model which performs a relevance based sequence modeling of the speech temporal sequence with short segment i-vectors for language recognition tasks using deep bidirectional long short term memory(BLSTM) neural networks. We refer to this as the i-BLSTM model throughout the rest of the paper.

The rest of the paper is organized as follows. In Section 2 we describe the proposed i-BLSTM model. Section 3 provides the details of experimental setup used for the LRE2017 and RATS datasets as well as the details of the state-of-the-art baseline i-vector model and the end-to-end LSTM based neural network approach to language recognition. The results of experiments for the two datasets and for various noisy environments are reported in Section 4 which is followed by a discussion on the i-BLSTM model in Section 5. In Section 6, we summarize the important contributions of this paper.

2. The hybrid i-BLSTM Model

The proposed i-BLSTM model is shown in Fig.1. The short-segment i-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

[*] Equal contribution to the work by the first two authors. This work was also performed when the first author was part of the LEAP lab, Indian Institute of Science.

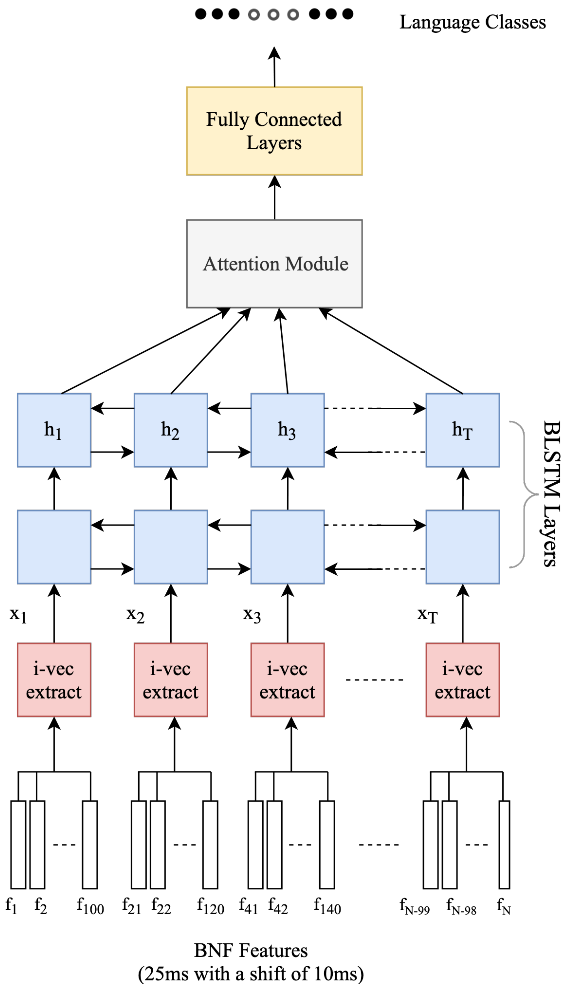

Figure 1: *Proposed i-BLSTM model for language recognition.*

bidirectional LSTM (BLSTM) [22] layers and attention module [13]. There are two BLSTM layers and the first BLSTM layer is fed with a variable length sequence of 500 dimensional short segment i-vectors. Both the forward and backward layers of the two BLSTM layers contain 256 cells each. Outputs from the forward and backward layers of BLSTM are concatenated before passing them onto the next layer and the output from the final BLSTM layer is fed to the attention module. The attention mechanism [13] shown in Fig. 2 provides an efficient way to aggregate the output sequence of final BLSTM layer. The model implements the following set of equations,

$$\mathbf{u}_t = tanh(\mathbf{W}_e \mathbf{h}_t + \mathbf{b}_e) \tag{1}$$

$$a_t = \frac{exp(\mathbf{u}_t^T \mathbf{u}_e)}{\sum_t exp(\mathbf{u}_t^T \mathbf{u}_e)} \tag{2}$$

$$\mathbf{e} = \sum_t a_t \mathbf{h}_t \tag{3}$$

The weights $\mathbf{W}_e$ and bias $\mathbf{b}_e$ of the attention module along with the vector $\mathbf{u}_e$ are learned in training. Normalized weights $\mathbf{a}_t$ are computed based on the similarity of the vectors $\mathbf{u}_t$ and $\mathbf{u}_e$, which are then used to output a fixed dimension embedding $\mathbf{e}$ of the input sequence. The embedding $\mathbf{e}$ is then mapped to the language targets through a layer of fully connected network and a softmax output layer.

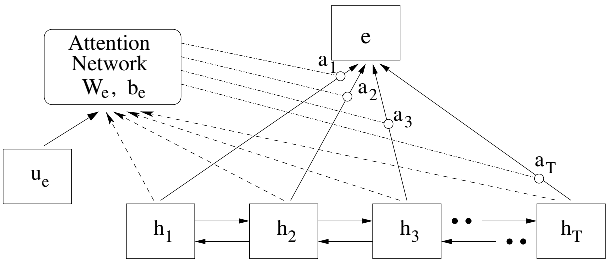

Figure 2: *Attention modeling in the proposed model*

3. Experimental Setup

3.1. Dataset

3.1.1. LRE2017 Dataset

LRE2017 setup has separate datasets for training, development and evaluation. All the models were trained on the LDC2017E22 training dataset and development dataset was used as a validation set for tuning the models. There are five major language clusters (Arabic, Chinese, English, Slavic and Iberian) with 14 target dialects. A total of 2069 hours spread across 16205 files is available for training. 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 3 sec, 10 sec or 30 sec duration and the audio extracted from video data consisting of 1000 sec recordings.

Further, we have added five different types of noise (Babble, Restaurant, Airport, Street, Subway) at various signal-to-noise ratios (0, 5, 10, 15 and 20 dB) to the LRE2017 evaluation dataset. The models were then evaluated on this noisy dataset to compare their relative robustness towards noise. Either the entire audio or only the initial half(to simulate non-stationary noise effects) is corrupted with noise. Also, to evaluate multi-speaker conditions, same language utterances from 2 different speakers of the LRE2017 evaluation dataset were merged. These scenarios were created to simulate the practical conditions where the audio can be corrupted with noise (stationary and non-stationary) and may contain multiple people speaking the same language. The proposed model and the i-vector baseline were also evaluated for their sensitivity towards absence of any speech activity detection (SAD) information.

3.1.2. RATS Dataset

The DARPA Robust Automatic Transcription of Speech (RATS) [3] 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 [3]. 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 [23].

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

Table 1: *Performance of reference and the developed system on the NIST LRE2017 evaluation dataset in terms of percentage accuracy, C_{avg} and EER.*

Dur./Model	LDA-SVM [21]	LSTM [28]	i-BLSTM
Accuracy (%)			
3	53.84	54.74	**54.80**
10	72.36	72.58	**75.89**
30	**82.98**	76.10	82.27
1000	**56.23**	42.86	54.07
overall	67.86	64.74	**68.65**
C_{avg}			
3	0.53	0.55	**0.50**
10	0.27	0.35	**0.26**
30	**0.13**	0.28	0.18
1000	0.54	0.79	**0.50**
overall	0.37	0.48	**0.36**
EER (%)			
3	**13.40**	15.39	15.47
10	6.47	8.70	**6.32**
30	**3.50**	7.25	3.67
1000	15.35	26.27	**14.71**
overall	**9.26**	14.38	9.65

purposes. All the models are trained as a 6 class classification problem. 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, 10 and 30 sec chunks of voiced data from the full length evaluation files.

3.2. Feature Extraction

The use of bottleneck features derived from a speech recognition acoustic model have recently shown consistent improvements for language recognition [24, 25]. We extract 80 dimensional bottleneck features (BNF) from a DNN trained for automatic speech recognition using Kaldi [26] framework. The model to extract bottleneck features was trained using 39 ($13 + \Delta + \Delta\Delta$) dimensional MFCC features with 10 msec frame rate over 25 msec windows on labeled speech data from Switchboard SWB1 and Fisher corpora ($\sim$2000 hours). The model uses 7 hidden layers with ReLU activation and layer-wise batch normalization. A speech activity detection (SAD) [27] algorithm was applied on the original audio to remove the unvoiced frames from the features extracted. Cepstral mean variance normalization (CMVN) over each utterance followed by a sliding window cepstral mean variance normalization (CMVN) over a 3 sec window was applied on the extracted features.

3.3. Baseline System

3.3.1. i-vector LDA-SVM

The i-vectors used in our experiments are features of fixed dimension extracted from a variable length sequence of bottleneck features (BNF) [7]. We follow the procedure described in [8] for their extraction.

Once extracted, the i-vectors are length normalized and their dimension reduced using linear discriminant analysis (LDA). Finally a Support Vector Machine (SVM) is trained on the i-vectors for language classification.

Table 2: *Performance of the systems in terms of accuracy (%) when evaluated on data corrupted partially and completely with noise at various SNR levels.*

SNR/Model	LDA-SVM [21]	LSTM [28]	i-BLSTM
No noise	72.36	72.1	**75.89**
Partially Noisy			
5dB	53.31	56.50	**59.79**
10dB	55.76	60.42	**63.02**
15dB	58.49	62.61	**65.90**
20dB	59.78	64.61	**68.16**
overall	56.83	61.03	**64.22**
Noisy			
5dB	47.93	48.36	**51.58**
10dB	53.77	56.30	**59.86**
15dB	57.82	61.63	**64.30**
20dB	60.00	65.28	**67.72**
overall	54.88	57.89	**60.87**

3.3.2. LSTM

Long Short Term Memory Recurrent Neural Networks (LSTM-RNN) based models were explored as end-to-end solutions for language recognition in [28, 29]. In [28], experiments on NIST datasets have shown that these models while better than the conventional i-vector based models on short duration (3 sec) test segments, fare poorly on longer duration test segments (10 sec, 30 sec). Their best performing LSTM model, which is a two layer LSTM with 512 units in each layer followed by an output softmax layer is implemented in this paper.

3.4. Performance Metrics

Since the LRE2017 and RATS are closed language set LID evaluations, we use the accuracy as the primary metric for evaluating various models in this work. We also report the official LRE cost metric $Cavg$ for reference on the original LRE evaluation set. It is worth noting that the cost metric of $Cavg$ can be improved with score calibration using a development set. In this paper, we have reported LID results from raw scores without any calibration.

4. Results

The performance of the proposed hybrid i-BLSTM model is compared with the baseline (LDA-SVM) and the LSTM model for the LRE2017 dataset (Table 1). In terms of the primary evaluation metric (C_{avg}), the proposed approach provides the best results for all duration conditions except the 30 sec condition. In addition, the proposed hybrid neural model improves significantly over the previous LSTM [28] based approach for longer duration conditions (10 sec or more). These results highlight that the proposed approach is effective in modeling long-term dependencies in the audio signal compared to previous neural network models for language recognition. We also find that the accuracy measure is well correlated with the C_{avg} measure. In the subsequent experiments on noisy datasets, we report the accuracy measure alone (all the trends found in accuracy measure for the noisy datasets are correlated with the C_{avg} results are well).

The performance on noisy and partial noisy versions of the LRE2017 dataset is shown in Table 2 for the 10 sec recordings. In the noisy and partial noisy conditions, the LSTM model [28]

Table 3: *Performance of the i-vector baseline system and the proposed model on the RATS dataset in terms of percentage accuracy, C_{avg} and EER.*

Dur./Models	LDA-SVM	i-BLSTM
Accuracy (%)		
3	65.39	**67.99**
10	77.46	**79.66**
30	85.38	**87.72**
full length	92.22	**92.50**
C_{avg}		
3	1.12	**0.89**
10	0.81	**0.65**
30	0.57	**0.44**
full length	0.40	**0.32**
EER (%)		
3	25.90	**21.48**
10	17.81	**14.00**
30	11.88	**8.98**
full length	7.64	**6.08**

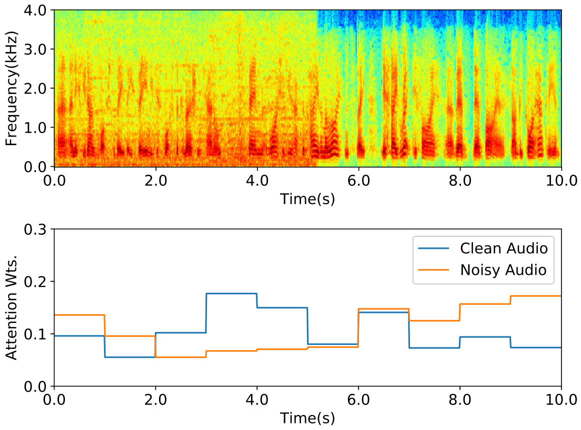

Figure 3: *Attention weights on partially noisy LRE2017 file and the corresponding clean LRE2017 recording.*

improves over the baseline i-vector SVM system. The proposed hybrid i-BLSTM further improves the performance on all the SNR conditions. On average, the proposed approach yields relative improvements of 14% over the baseline system for the partially noisy condition and about 11% for the noisy condition.

On the RATS dataset, the language recognition results are reported in Table 3. The RATS recordings are inherently noisy due to the transmission artifacts. On all the duration conditions (and in terms of all the performance metrics considered), the i-BLSTM model improves over the baseline system. The improvements are more pronounced in short duration recording conditions. The proposed i-BLSTM model is more robust to the channel artifacts compared the baseline i-vector SVM model.

5. Discussion

5.1. Attention Analysis

In this subsection, we analyze the role of the attention mechanism in the proposed i-BLSTM model. We plot the spectrogram of a partially corrupted speech recording (first 5 sec at 10 dB SNR) and the corresponding attention vector which is computed at 1 sec resolution in Fig. 3. The attention weights for the same recording without any additive noise (clean recording) is also shown for reference. As can be seen, for the partially noisy recording, the attention weights for the later part of the utterance are relatively higher making them more relevant to the task. Comparing the attention weights for the clean and noisy recording reveals that the attention mechanism, which gave high relevance to the early part of the audio in the clean conditions, is dynamic to the change in the SNR conditions of the audio. The model is able to shift the focus to the regions of the audio that are more informative for the task of language recognition.

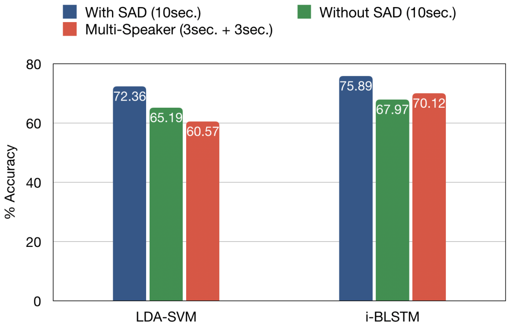

Figure 4: *Performance comparison for the baseline system and i-BLSTM model for multi-talker LID and LID without SAD.*

5.2. Multi-talker LID and LID without SAD

We also performed two additional LID experiments with the proposed i-BLSTM model. The first experiment used speech recordings in testing that contain multiple speakers. This is obtained by merging 3 sec speech utterances in the clean LRE evaluation set from multiple talkers of the same language. The second experiment explores the sensitivity of the LID systems to the absence of any speech activity detection (SAD) information on the 10 sec recordings. As seen in Fig. 4, the proposed model is more robust to the presence of multiple talkers in the evaluation dataset. Also, the baseline i-vector LDA-SVM model experiences a significant drop in performance in the absence of SAD information while the i-BLSTM model is relatively less sensitive. These experiments confirm that the hybrid i-BLSTM framework is able to efficiently model the time series for the language classification by relevance weighting based on the attention mechanism.

6. Summary

In this paper, we have proposed a novel approach for language recognition using a hybrid i-BLSTM model. The input to the model is the sequence of i-vector features extracted using 1 sec windows and the model architecture contains a bi-directional LSTM with attention. Several experiments on the language recognition task for LRE2017 and RATS dataset highlight the significant improvements obtained by the proposed model. The additional analysis using noisy data and language recognition experiments in the absence of speech activity detection shows that the attention mechanism is effective in dynamically reweighting the 1 sec i-vector segments based on their relevance to the language classification task.

7. References

[1] A. Waibel, P. Geutner, L. M. Tomokiyo, T. Schultz, and M. Woszczyna, "Multilinguality in speech and spoken language systems," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1297–1313, 2000.

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

[3] K. Walker and S. Strassel, "The rats radio traffic collection system," in *Odyssey 2012-The Speaker and Language Recognition Workshop*, 2012.

[4] H. Li, B. Ma, and K. A. Lee, "Spoken language recognition: from fundamentals to practice," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1136–1159, 2013.

[5] S. Ganapathy, M. Omar, and J. 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.

[6] M. McLaren, D. Castan, and L. 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.

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

[8] N. Dehak, P. A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language recognition via i-vectors and dimensionality reduction," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

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

[10] D. Snyder, D. Garcia-Romero, and D. 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.

[11] S. Ganapathy, K. Han, S. Thomas, M. Omar, M. V. Segbroeck, and S. S. Narayanan, "Robust language identification using convolutional neural network features," in *Fifteenth annual conference of the international speech communication association*, 2014.

[12] B. Padi, S. Ramoji, V. Yeruva, S. Kumar, and S. 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.

[13] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[14] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[15] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[16] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. 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.

[17] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. 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.

[18] B. Padi, A. Mohan, and S. Ganapathy, "End-to-end language recognition using attention based hierarchical gated recurrent unit models," in *Proc. ICASSP*, 2019.

[19] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2227–2231.

[20] C.-W. Huang and S. S. Narayanan, "Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition," in *Multimedia and Expo (ICME), 2017 IEEE International Conference on*. IEEE, 2017, pp. 583–588.

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

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

[23] 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.

[24] F. Richardson, D. Reynolds, and N. Dehak, "Deep neural network approaches to speaker and language recognition," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1671–1675, 2015.

[25] Y. Lei, N. Scheffer, L. Ferrer, and M. 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.

[26] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[27] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, pp. 1–3, 1999.

[28] R. Zazo, A. Lozano-Diez, and J. 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.

[29] J. Gonzalez-Dominguez, I. Lopez-Moreno, H. Sak, J. González-Rodríguez, and P. J. Moreno, "Automatic language identification using long short-term memory recurrent neural networks," in *INTERSPEECH*, 2014.