International Journal of Speech Technology (2018) 21:501–508
https://doi.org/10.1007/s10772-017-9481-6

Combining evidences from excitation source and vocal tract system features for Indian language identification using deep neural networks

Mounika Kamsali Veera[1] · **Ravi Kumar Vuddagiri**[1] · **Suryakanth V. Gangashetty**[1] · **Anil Kumar Vuppala**[1]

Received: 15 July 2017 / Accepted: 10 November 2017 / Published online: 12 December 2017

Abstract
In this paper, a combination of excitation source information and vocal tract system information is explored for the task of language identification (LID). The excitation source information is represented by features extracted from linear prediction (LP) residual signal called the residual cepstral coefficients (RCC). Vocal tract system information is represented by the mel frequency cepstral coefficients (MFCC). In order to incorporate additional temporal information, shifted delta cepstra (SDC) are computed. An LID system is built using SDC over both MFCC and RCC features individually and evaluated based on their equal error rate (EER). Experiments have been performed on a dataset consisting of 13 Indian languages with about 115 h for training and 30 h for testing using a deep neural network (DNN), DNN with attention (DNN-WA) and a state-of-the-art i-vector system. DNN-WA outperforms the baseline i-vector system. An EER of 9.93 and 6.25% are achieved using RCC and MFCC features respectively. By combining evidence from both features using a late fusion mechanism, an EER of 5.76% is obtained. This result indicates the complementary nature of the excitation source information to that of the widely used vocal tract system information for the task of LID.

Keywords Language identification · Shifted delta cepstral features · Deep neural network · Attention mechanism · I-vector

1 Introduction

Automatic language identification (LID) refers to the task of identifying the language of a spoken utterance Muthusamy et al. (1994). LID has a wide range of applications in a multilingual speech recognition system as a front-end in information services (customer care) and in providing many computer and telephone based services. Any spoken utterance contains information about the speaker, emotion of the speaker, channel, environment and other variable factors. The presence of such variability makes it a challenge to identify the language spoken in a given utterance invariant to such factors. Especially, in an Indian scenario, where almost every state has a language of its own and every language has many dialects, the task of identifying a language becomes more difficult. In the current work, we consider 13 official Indian languages.

Speech is produced when the vocal tract system is excited by an excitation source. Any spoken utterance would thus contain excitation source as well as the vocal tract system information. The vocal tract information is mainly represented by vocal tract system features like the mel frequency cepstral coefficients (MFCC) and linear prediction cepstral coefficients (LPCC). Excitation source information can be represented by residual cepstral coefficients (RCC) extracted from the LP residual of the speech signal. MFCCs are the widely used spectral features for the LID task. They represent the gross characteristics of the vocal tract system. Prosodic features have been explored for capturing the language information from an utterance in Mary et al. (2005). In Mary and Yegnanarayana (2008) prosodic features are extracted by considering the syllable as the basic unit for

✉ Mounika Kamsali Veera
mounika.kv@research.iiit.ac.in

Ravi Kumar Vuddagiri
ravikumar.v@research.iiit.ac.in

Suryakanth V. Gangashetty
svg@iiit.ac.in

Anil Kumar Vuppala
anil.vuppala@iiit.ac.in

[1] Speech Processing Laboratory, International Institute of Information Technology, Hyderabad, India

extracting language-specific information. In Rao and Nandi (2015), the usefulness of excitation source information is explored for the task of LID. Excitation source information has been explored for speaker recognition (SR) Pati and Prasana (2011), emotion recognition Al-Talabani et al. (2013), speech enhancement Yegnanarayana et al. (2002) tasks. In this work, a late fusion mechanism is used to combine the information from source and system level features for the LID task.

A Gaussian mixture model (GMM) based classification has been the most successful approach for LID Torres-Carrasquillo et al. (2002). The current state-of-the-art i-vector based LID system also uses the universal background model (UBM) as a front-end and then a classifier is trained over the model. i-vector system has been a successful approach for both SR and LID. The two tasks being closely related, most of the methods applied for SR will apply equally well for LID Richardson et al. (2015), Dehak et al. (2011). Building an i-vector based LID system requires high computation and since the feature extraction and classification are done separately, it is very time consuming.

DNNs have been recently proposed for the task of LID. The profound performance of DNN in acoustic modeling for the task of speech recognition has motivated many speech researchers to explore the DNNs for different tasks Bengio (2009). While previous attempts to use neural networks for LID task used shallow architectures, in this study, we use DNN. With the large amount of training data, DNNs have proven to be successful acoustic models for speech recognition Hinton et al. (2012). In Richardson et al. (2015), a single DNN was trained for the task of LID and SR, and significant gains have been reported for both the tasks. In Lopez-Moreno et al. (2014), authors reported the DNN performance to be superior to that of the state-of-the-art i-vector system. DNN was either user as an acoustic model or as a classifier so far, but in Lakhsmi et al. (2016), an end-to-end DNN was built for the SR task. Similarly, in our current work, an end-to-end DNN and a modified DNN architecture called with attention mechanism (DNN-WA) Raffel and Ellis (2015) is used to address the task of LID.

In Nandi et al. (2014), the complementary nature of RCC to that of the MFCC information for the task of LID is demonstrated. Current work is similar in spirit with the hypothesis of Nandi et al. (2014) that the excitation source and vocal tract system have significant contributions for producing the sound units of a particular language. While Nandi et al. (2014) used the basic GMM for classification, in this work, we use an end-to-end DNN-WA mechanism to validate the hypothesis in a DNN framework.

Rest of the paper is organized as follows: In Sect. 2, a detailed description of the Indian language database used for our experiments is given. In Sect. 3, generation of feature vector from excitation source and vocal tract system using SDC features is given, followed by the description of the DNN-WA mechanism in Sect. 4. The experiments and results are detailed in Sect. 5. Finally, the summary and conclusions are presented in Sect. 6.

2 Indian language database

The Indian language database consists of 13 of the official Indian languages.

The speech data is collected in read speech mode at 16 kHz sampling rate. Every language has significant amount of male and female speakers to account for the gender and speaker variability. In our experiments, each speech file is sliced into 5 s chunk both in the training and testing datasets. The details of number of speakers (#male, #female) per language and the amount of data available (#hours) per language is given in Table 1. Most of the Indian languages originate from the Sanskrit language leading to a common set of phonemes and similar grammatical structure. This similarity makes the development of an LID system challenging.

3 Shifted delta cepstral (SDC) features

The most widely used features for the task of LID are MFCC. Dynamic information from MFCC could be obtained by including static + delta + delta delta features. Computing the delta cepstra in this way captures temporal information over fewer frames alone, i.e., less context. An improved LID performance could be obtained using SDC feature vectors created by stacking delta cepstra compounded across multiple speech frames Torres-Carrasquillo (2002) thus capturing the context to a greater extent. The computation of SDC is illustrated in Fig. 1.

SDC construction requires stacking of the dynamic features computed over few frames along with the static features. The SDC features are specified by four parameters, N, d, P, k, where N is the number of cepstral coefficients considered from each frame, d represents the time delay and time advance for delta computation. P represents the time shift between consecutive delta computed blocks and k represents the number of blocks whose delta coefficients are concatenated together, where i = 0, 1, 2, ...k−1.

$$\Delta c(t) = c(t + iP + d) - c(t + iP - d) \tag{1}$$

Given a spoken utterance, SDC feature vector is computed for every frame time t considering the context across P × k frames based on the configuration. In this work, we use the standard SDC configuration used for MFCC features Torres-Carrasquillo (2002) along with some expreimentation in finding optimal configuration for computing SDC over the RCC. From the trials, it was found that 10-1-3-3

Table 1 Description of the Indian language database used Mounika et al. (2016)

Language	Train			Test		
	#Hours	#M	#F	#Hours	#M	#F
Assamese	11.79	19	10	1.94	3	3
Bengali	9.54	24	35	1.36	15	14
Gujarati	9.64	115	75	2.17	35	36
Hindi	10.03	40	26	3.16	15	17
Kannada	10.04	21	16	0.97	10	4
Malayalam	7.04	7	6	2.86	9	7
Manipuri	3.81	5	6	1.45	3	2
Marathi	7.82	72	31	2.47	17	15
Odiya	8.55	31	30	2.18	9	9
Punjabi	14.06	2	9	3.74	2	1
Tamil	4.07	12	8	0.87	5	4
Telugu	10.30	57	21	3.08	4	4
Urdu	7.63	55	18	3.03	16	5

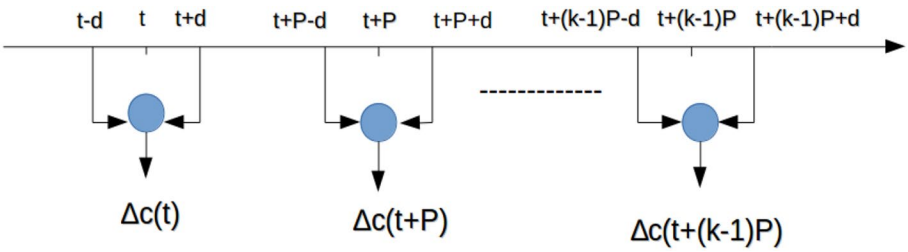

Fig. 1 Computation of SDC feature vector at frame t for configuration N-d-P-k

configuration works better for RCC features than the popular 7-1-3-7 configuration that is widely used for MFCCs.

4 Deep neural network with attention (DNN-WA)

Language information mostly exists in longer durations. Information from shorter durations (sub-segmental information) corresponds to excitation cues. We use the RCC features extracted from the LP Residual signal to represent the excitation source information and MFCC features to represent the vocal tract system features. When DNN based classifier is used to model short durational cues, it fails to capture the long-term dependencies that are reqired for LID. To use the complementary language discriminative information from excitation source, the modeling mechanism should be capable of modeling the long-term dependencies. A regular DNN is a frame based classifier where the decision from every frame is given equal importance. The average over all frames would represent the final decision to the utterance. Unlike DNN, DNN-WA is an utterance based decision mechanism which accounts for the long-term relations by implicitly making use of the context within the utterance.

The architecture of DNN-WA is shown in Fig. 2. The attention mechanism in a DNN helps in visualizing how the

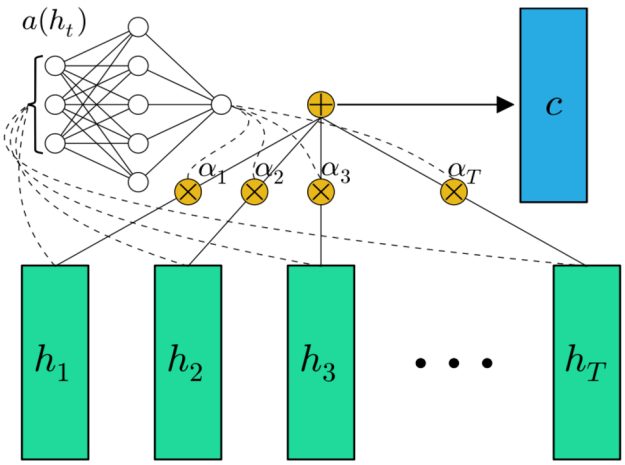

Fig. 2 DNN-WA Raffel and Ellis (2015)

input is being attended while taking the decision. In Bahdanau et al. (2014), attention mechanism was explored for neural machine translation. In our work, we use the DNN-WA for the task of LID. In Bahdanau et al. (2014), input as well as the output features were used for attention, but in current work, we use the input feature vector alone to do so.

Given an input sequence, $X = \{x_1, x_2, \ldots, x_T\}$, T represents the number of frames in an utterance. Based on the

architecture of the DNN, a hidden layer representation, $H = \{h_1, h_2, \ldots, h_T\}$, is computed by forward propagation. The attention mechanism $a(h_t)$ is represented using a single layer perceptron which learns the weight to be given for each hidden layer decision towards the final decision taken at utterance level. The output layer is a *softmax* function that normalizes the values between zero and one.

$$\begin{aligned} H &= [h_1, h_2, \ldots h_T] \\ \gamma &= \tanh(W_a H + b_a) \\ \alpha &= softmax(\gamma) \end{aligned} \tag{2}$$

In the above equations, α is referred to as *attention vector*, and W_a, b_a are the parameters of the attention network optimized along with other parameters of the network using back-propagation algorithm.

The context vector is computed from the attention vector as

$$c = H\alpha \tag{3}$$

The output is computed by transforming the context vector c using output layer weights U followed by *softmax* operation.

$$y = softmax(Uc + b_o) \tag{4}$$

where b_o is the output layer bias. Note that for the entire input utterance X only a single decision vector y is predicted.

Given a spoken utterance with T frames a single decision y is computed at the utterance level.

5 Experiments and results

In this section a brief introduction to the state-of-the-art i-vector system is given and later the proposed DNN-WA is used for an utterance level classification for the LID task.

The State-of-the-art i-vector based LID system was first introduced in Dehak et al. (2011). The i-vector based system operates on the assumption that all the pertinent variability is captured by a low rank rectangular matrix T named the total variability matrix. The GMM supervector (vector created by stacking all mean vectors from the GMM) for a given utterance can be modeled as given in Eq. 5

$$M = m + Tw + \epsilon \tag{5}$$

where m is the UBM supervector, w is the i-vector and ϵ is the residual noise term. Conceptually, an i-vector captures the sequence summary given an utterance but is computationally intensive. Once the i-vector is computed, a dimension reduction technique like Principal Components Analysis (PCA) is used to reduce the length of the sequence vector (context summarization). In this work we propose to use a DNN-WA model that generates a context vector that is similar to an i-vector in terms of sequence summarization. While other models like Recurrent Neural Networks (RNN) help serve the purpose, they are not parellelizable and require more computational power. Hence, we propose to use a DNN-WA mechanism.

In this work, we use 2048 mixtures for UBM building and the dimension of the T matrix is set to 100. In Table 2, a comparision with the state-of-the-art i-vector system has been made. We compare our results as i-vector vs DNN vs DNN-WA. Although there is little improvement achieved in moving from i-vector to DNN, a significant improvement is observed with DNN-WA. This improvement comes from the capability of the DNN-WA mechanism to summarize the entire utterance into a single context vector that further helps in giving an utterance level decision.

A set of experiments have been conducted to select the best dimension and type of feature i.e., static vs dynamic vs SDC. SDC outperformes the other features and hence we proceed with the SDC features applied on MFCC as well as RCC. From the literature, the SDC configuration 7-1-3-7 is choosen for MFCCs. Further, we found that 10-1-3-3 SDC configuration to work better for RCC. Thus 56D MFCC based and 40D RCC based LID systems are considered for late fusion using DNN-WA.

In this section results are reported for 3 systems MFCC, RCC and MFCC + RCC based LID system across different dimensions. The results given in Tables 2, 3, 4 are obtained

Table 2 Equal error rate (EER in %) of MFCC based LID system

Language	Ass.	Ben.	Guj.	Hin.	Kan.	Mal.	Man.	Mar.	Odi.	Pun.	Tam.	Tel.	Urd.	Average
$i-vector_{13D}$	6.05	14.1	18.1	26.1	16.2	10.6	5.25	13.8	13	13.5	26.9	5.79	6.94	13.59
$i-vector_{39D}$	5.12	13	13.2	19.9	14.6	7.51	4.40	10.1	6.49	7.16	19.9	4.99	5.73	10.18
DNN_{13D}	3.78	7.59	10.71	25	16	23.97	7.18	15.40	7.53	5.60	27.78	3.72	8.06	12.49
DNN_{39D}	4.30	6.59	8.2	22	15.2	18.61	5.6	12.2	4.62	4.99	20	3.04	6.6	10.15
$DNN-WA_{39D}$	6.10	6.32	6.56	25.4	7.8	15.2	4.8	7.4	5.84	4.8	8.8	4.6	6.57	8.48
DNN_{56D}	8.97	3.97	8.68	6.91	5.71	5.93	5.6	12.31	5.96	6.2	11.29	2.91	6.83	7.02
$DNN-WA_{56D}$	6.32	6.21	7.10	6.41	5.78	5.83	6.33	7.1	6.29	5.2	6.82	5.36	6.43	**6.25**

Bold values indicate the average equal error rate of all the target languages considered

Table 3 Equal error rate (EER in %) of RCC based LID system

Language	Ass.	Ben.	Guj.	Hin.	Kan.	Mal.	Man.	Mar.	Odi.	Pun.	Tam.	Tel.	Urd.	Average
DNN_{14D}	0.726	12.06	12.61	32.4	13.15	30.18	17.6	18.19	10.18	6.741	44.41	4.85	16.82	16.92
DNN_{40D}	1.66	9.40	9.47	28.60	9.40	18.70	14.93	15.6	8.81	5.42	34.89	5.10	19.69	13.97
DNN_{56D}	2.45	7.25	9.4	36.2	10	14.51	12.14	15.29	9.88	4.64	29.34	5.06	19.4	13.50
$DNN - WA_{14D}$	5.05	7.2	7.4	24.6	10.6	19.6	5.90	8.47	9.2	5	19.4	5	6	10.27
$DNN - WA_{56D}$	7.2	7.6	7.83	44.1	12.8	16.2	8.01	9.2	7.30	6.75	27.4	5	7.05	12.81
$DNN - WA_{40D}$	7.22	7	7.51	24.8	9.41	14.1	4.6	8.8	5.79	5.4	23.9	4.28	6.19	**9.93**

Bold values indicate the average equal error rate of all the target languages considered

Table 4 Equal error rate (EER in %) of the proposed LID system built by integration of MFCC and RCC based LID systems

Language	Ass.	Ben.	Guj.	Hin.	Kan.	Mal.	Man.	Mar.	Odi.	Pun.	Tam.	Tel.	Urd.	Average
$DNN - WA_{MFCC-SDC}$	6.32	6.21	7.10	6.41	5.78	5.83	6.33	7.1	6.29	5.2	6.82	5.36	6.43	6.25
$DNN - WA_{RCC-SDC}$	7.22	7	7.51	24.8	9.41	14.1	4.6	8.8	5.79	5.4	23.9	4.28	6.19	9.93
$DNN - WA_{combinedevidence}$	5.68	6.4	5.76	7.28	6.02	6.73	4.4	6.32	4.96	4.81	7.57	3.71	5.26	**5.76**

Bold values indicate the average equal error rate of all the target languages considered

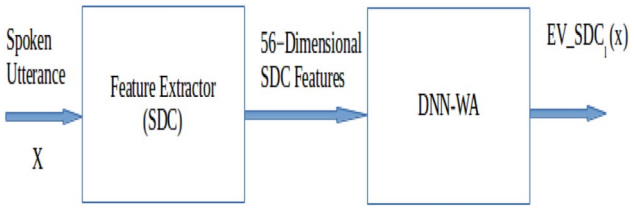

Fig. 3 Block diagram of the LID system using MFCC features

over a testing set consisting of 500 utterances per language each of 5 s duration. For every language 500 utterances were held out for validation purpose from the training data.

5.1 LID system built using SDC over vocal tract features (MFCC-SDC)

The block diagram of LID system built using MFCC-SDC features is shown in Fig. 3. Given a spoken utterance, 13 dimensional MFCCs (static) are extracted from each frame of 20 ms with 10 ms frame shift. The configuration 7-1-3-7 has been used to compute MFCC-SDC feature vector. Resulting 49 dimensional vector is stacked with the current frame's static features (7 dimensional) making it a 56 dimensional feature vector. DNNs were trained using mini-batch stochastic gradient descent with classical momentum. The size of mini-batch was equal to the length of the sequence given as input to the network. Normalized initialization is used for intializing the network. The dataset used for the experiments consists of 115 and 30 h of data for training and testing respectively. Validation set was used for hyper-parameter tuning. DNN was trained using mini-batch SGD, where the size was equal to the length of the sequence given as input to the network.

The input and output layers of the DNN and DNN-WA are 56 dimensional and 13 dimensional respectively. Hidden layer dimensions are variable and rectified linear units (ReLU) units Zeiler et al. (2013) were used as the hidden layer activation function. While regular DNN is frame based, the average over all frames of an utterance represents the final decision. In DNN-WA, the attention mechanism helps in direct utterance level classification. EER per language is used as the performance metric.

Experiments were performed to determine the depth and breadth of the network for the LID task.The network comprises of four hidden layers and each layer comprising of 700, 500, 200 and 100 units with ReLU activation functions and output is a soft-max layer with a cross-entropy cost function. Table 2 describes the results obtained with different architectures. The first column of the Table 2 represents the model used with the subscript representing the dimension of the input. Static, dynamic (static + delta + delta delta), and MFCC-SDC features are represented as 13D, 39D and 56D respectively. Column 2–13 represent the EER of the corresponding languages. Final column represents the average EER. Rows 2 and 3 represent the state-of-the-art i-vector based LID results. Row 4 onwards are the DNN based LID results. From Table 2 it is clear that the performance of DNN-WA is superior to the regular frame-based DNN model. Hence, we use the end-to-end DNN-WA mechanism in our further experiments.

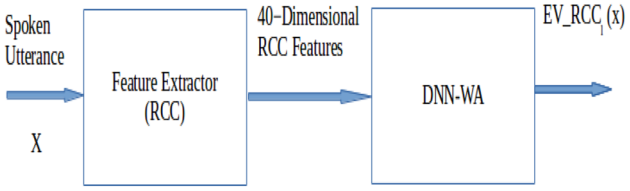

Fig. 4 Block diagram of the LID system using RCC features

5.2 LID system built using SDC over excitation source features (RCC-SDC)

Given a speech signal, LP residual is the error obtained when the signal is processed using tenth-order LP analysis. LP analysis removes the second-order relations from the speech signal. The higher order relations among the samples of the speech signal remain in the residual signal. The LP residual mostly contains the information of the excitation source. RCC features are used to represent such information about the excitation source.

The block diagram for LID system built using RCC-SDC features is shown in Fig. 4. We use 14 dimensional RCC features coming from each frame of 20 ms with a frame shift of 10 ms. The configuration 10-1-3-3 has been used to compute SDC feature vector. This configuration has been choosen based on experiments. SDC obtained as a 30 dimensional vector is then stacked with the current frame's static features (10 dimensional) making it a 40 dimensional feature vector. An end-to-end DNN-WA is used for utterance level classification. DNN-WA mechanism shows a significant improvement in performance compared to the regular DNN. DNN-WA has the ability to make use of the context by modeling the long-term dependencies within the utterance. The result of a 4-layer DNN-WA using RCC-SDC features is given in Table 3 with an EER of 9.9367%.

Table 3 describes the results obtained with different architectures. The first column of the Table 3 represents the model used with the subscript representing the dimension of the input. Static and RCC-SDC features are represented as 13D and 40D, 56D. Based on the choice of configuration for SDC computation, 40D and 56D come from 10-1-3-3 and 7-1-3-7 respectively. Columns 2–13 represent the EER of the corresponding languages. Final column represents the average EER. Rows 2 and 3 represent the state-of-the-art i-vector based LID results. Row 4 onwards are the DNN based LID results. From Table 3 it is clear that the DNN-WA outperforms the regular frame-based DNN model for an RCC based LID system.

In the LP residual signal, the region around the glottal closure instant (GCI) within each pitch period corresponds to high signal-to-noise-ratio called the sonority regions. These regions tend to be more robust to noise and are of great use in identifying a language Vuppala et al. (2015). This could mean that the sonority regions play a key role in decision making given an utterance. Plotting a speech signal spectrum and the attention vector weights helps in visualizing which frame or region of the signal has a major role in decision making. Interestingly, from Fig. 5 it is the transition regions in an utterance that contribute heavily towards identifying a language from a spoken utterance.

5.3 LID system built using combined evidence from MFCC-SDC and RCC-SDC features

Block diagram of the LID system built by combining the MFCC-SDC and RCC-SDC based systems is shown in Fig. 6. The result of combined evidence from excitation source and vocal tract system features is given in Table 4. The DNN scores obtained using the MFCC-SDC based LID system and the RCC-SDC based LID system are combined using late fusion mechanism, where the sum rule is used for integration. In the decision logic, the language label associated with the highest combined evidence is hypothesised as the language $L(\mathbf{x})$ of the test utterance x. That is,

$$L(\mathbf{x}) = \arg\max_{l} \{\log(EM_l(\mathbf{x})) + \log(ER_l(\mathbf{x}))\} \quad (6)$$

where $EM_l(\mathbf{x})$ and $ER_l(\mathbf{x})$ corresponds to the normalised DNN evidence for language l obtained using MFCC and RCC features respectively, for the input test utterance x.

EER is used as the performance measure. The EER for LID system built using the MFCC-SDC was 6.25% and LID

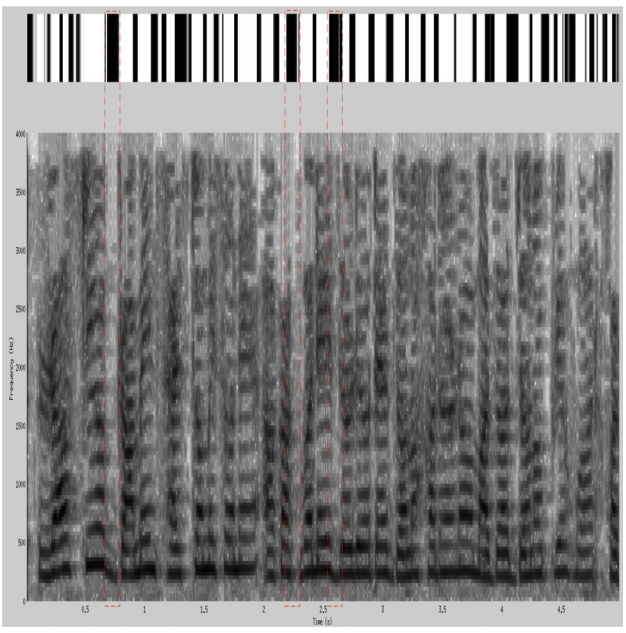

Fig. 5 A sample plot of an utterance with the attention values for respective frames in the utterance

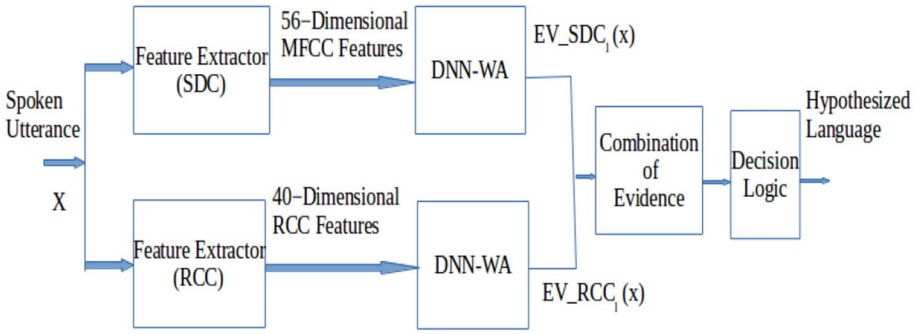

Fig. 6 Block diagram of the proposed LID system built by integration of MFCC and RCC features

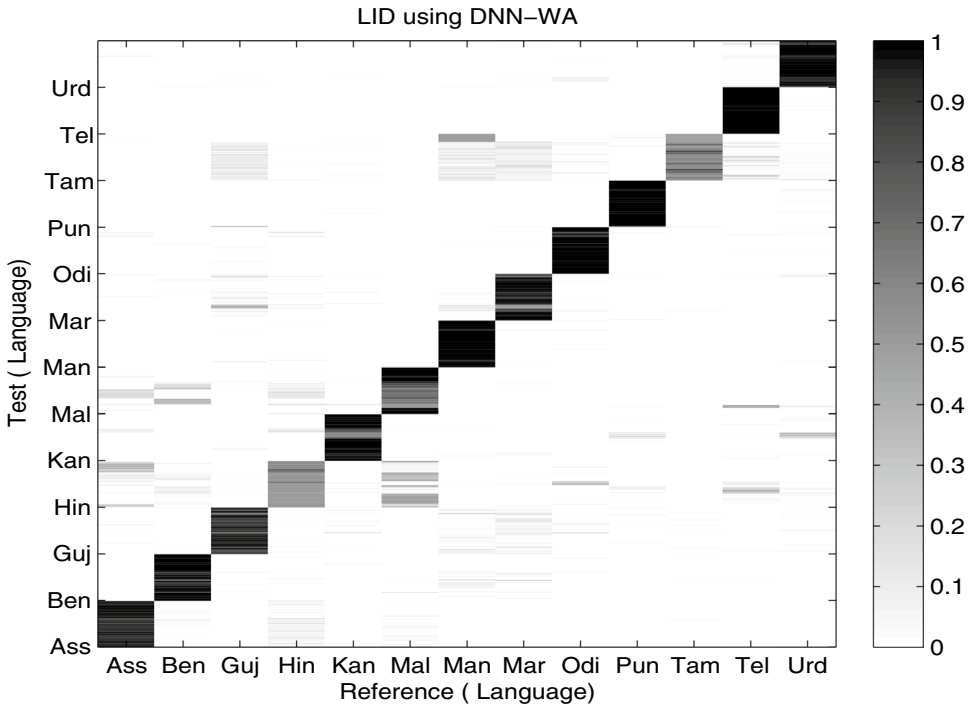

Fig. 7 Confusion matrix for DNN-WA model after late fusion

system built using RCC-SDC obtained 9.93%. The combined evidence used for LID resulted in an EER of 5.76%. A plot of confusion matrix obtained using the combined evidence by considering 500 utterances per language is shown in Fig. 7.

6 Summary and conclusion

In this paper, excitation source information using the RCC features and the vocal tract system information using the MFCC features are used for the LID task. Recently proposed DNN-WA mechanism is used for an utterance level classification. By combining the evidences from RCC based LID system with the conventional MFCC based LID system, an improved EER of 5.76% is achieved, which is better than the individual EER of 9.93 and 6.25% respectively. This result shows that the source information indeed helps in improving the performance of the LID by providing language information that is complementary to the vocal tract system information. Confusion matrix indicates greater difficulty in identifying the languages Hindi and Malayalam compared to others. As a part of future work, other DNN architectures that can deal with such challenges efficiently can be explored. New activation functions can also be tried.

Acknowledgements The authors would like to thank Science & Engineering Research Board (SERB) for funding "Language Identification in Practical Environments (YSS/2014/000933)" project.

References

Al-Talabani, A., Sellahewa, H., & Jassim, S. (2013). Excitation source and low level descriptor features fusion for emotion recognition using SVM and ANN. In Proceedings of Computer Science and Electronic Engineering Conference (CEEC) (pp. 156–161). IEEE.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. In Proceedings of International Conference on Learning Representations (ICLR).

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, *2*(1), 1–127.

Dehak, N., Torres-Carrasquillo, P. A., Reynolds, D. A., & Dehak, R. (2011). Language recognition via i-vectors and dimensionality reduction. In INTERSPEECH (pp. 857–860).

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, Ar, Jaitly, N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, *29*(6), 82–97.

Lakhsmi, H. R., Achanta, S., Bhavya, P. V., & Gangashetty, S. V. (2016). An investigation of end-to-end speaker recognition using deep neural networks. *International Journal of Engineering Research in Electronic and Communication Engineering*, *3*(1), 42–47.

Leena, M., Rao, K. S., & Yegnanarayana, B. (2005). Neural network classifiers for language identification using phonotactic and prosodic features. In Proceedings of International Conference on Intelligent Sensing and Information Processing (pp. 404–408). IEEE.

Lopez-Moreno, I., Gonzalez-Dominguez, J., Plchot, O., Martinez, D., Gonzalez-Rodriguez, J., & Moreno, P. (2014). Automatic language identification using deep neural networks. In Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (pp. 5337–5341). IEEE.

Mary, L., & Yegnanarayana, B. (2008). Extraction and representation of prosodic features for language and speaker recognition. *Speech Communication*, *50*(10), 782–796.

Mounika, K. V., Achanta, S., Lakshmi, H. R., Gangashetty, S. V., & Kumar Vuppala, A. (2016). An investigation of deep neural network architectures for language recognition in Indian languages. In INTERSPEECH (pp. 2930–2933).

Muthusamy, Y. K., Barnard, E., & Cole, R. A. (1994). Reviewing automatic language identification. *IEEE Signal Processing Magazine*, *11*(4), 33–41. https://doi.org/10.1109/79.317925.

Nandi, D., Pati, D., & Rao, K. S. (2014). Sub-segmental, segmental and supra-segmental analysis of linear prediction residual signal for language identification. In Proceedings of International Conference on Signal Processing and Communications (SPCOM) (pp. 1–6). IEEE.

Pati, D., & Prasana, S. R. M. (2011). Subsegmental, segmental and suprasegmental processing of linear prediction residual for speaker information. *International Journal of Speech Technology*, *14*(1), 49–64.

Raffel, C., & Ellis, D. P. W. (2015). Feed-forward networks with attention can solve some long-term memory problems. http://arxiv.org/abs/1512.08756.

Rao, K. S., & Nandi, D. (2015). Implicit excitation source features for language identification. In Language Identification Using Excitation Source Features (pp. 31–51). New York: Springer.

Richardson, F., Reynolds, D., & Dehak, N. (2015). A unified deep neural network for speaker and language recognition. In INTERSPEECH (pp. 1146–1150).

Torres-Carrasquillo, P. A., et al. (2002). Approaches to language identification using gaussian mixture models and shifted delta cepstral features. In INTERSPEECH.

Torres-Carrasquillo, P. A., Singer, E., Kohler, M. A., Greene, R. J., Reynolds, D. A., & Deller, J. R., Jr. (2002). Approaches to language identification using Gaussian mixture models and shifted delta cepstral features. In INTERSPEECH.

Vuppala, A. K., Mounika, K. V., & Vydana, H. K. (2015). Significance of speech enhancement and sonorant regions of speech for robust language identification. In Proceedings of Signal Processing, Informatics, Communication and Energy Systems (SPICES) (pp. 1–5). IEEE.

Yegnanarayana, B., Prasana, S. R. M., & Rao, K. S. (2002). Speech enhancement using excitation source information. In Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (vol. 1, pp. I–541). IEEE.

Zeiler, M., Ranzato, M., Monga, R., Mao, M., Yang, K., Le, Q., Nguyen, P., Senior, A., Vanhoucke, V., Dean, J., & Hinton, G. (2013). On rectified linear units for speech processing. In Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (pp. 3517–3521). IEEE.