Decision Level Fusion based Approach for Indian Languages Identification using Deep Neural Network

Kanika Gupta
speech processing lab
International Instiute of information technology
Hyderabad,India
kanika.gupta@research.iiit.ac.in

Kartikeya Singh Gour
School of engineering
Central university of Karnataka
Kalaburagi,Karnataka,India
singh.kartikeya16@gmail.com

Sompal Arya
School of engineering
Central university of Karnataka
Kalaburagi,Karnataka,India
sompalsrya99@gmail.com

Suryakanth V. Gangashetty
speech processing lab
International Instiute of information technology
Hyderabad,India
svg@iiit.ac.in

***Abstract*—In this work, we explore the combination of MFCC and RCC features for Indian Language identification(LID) task. MFCC represents the vocal tract information while RCC represents the excitation source information, which are complementary to each other. We use Deep neural network (DNN) framework for end to end learning. To combine these two information we uses different fusion techniques such as Min, Max, Average and Min-Max for decision level fusion. We perform experiments on 13 Indian languages data set. We compare our results with state of art I-vector based approach. Min-max based fusion approach gives 9.64% ERR compare to 10.18% of 39D I-vector based approach.**

Keywords—Deep neural network, MFCC, RCC, Fusion

I. Introduction

To identify the language of spoken word from an unknown speaker called as Automatic Language identification(LID). LID have multiple applications such as multilingual chatbots, automatic answer systems for voice calls and speech based personal assistant in mobile phones. Number of languages, speakers and length of sentences makes LID a challenging problems. LID depends on speaker emotions, enviorments and other variable factors also. In india every state has its own language and every language has multiple dialetcts which makes LID task more challenging. In this work we use 13 official indian languages for experiments.

An spoken utterance contain a lot of information which we can use to build LID. Vocal tract plays an important role in speech production. When vocal tract system is excited by excitaion source(air) then speech produces. A spoken world decompose two parts of information, one is vocal tract information and second is excitation source information. Vocal tract system mainly represented by the mel frequency cepstral coefficients(MFCC) features and excitation source information mainly represented by residual cepstral coefficients(RCC). In this work to represents the full speech production system information we combine the information present in MFCC and RCC using feature level fusion approach.

MFCC is the most appropriate features to represent the vocal tract system. MFCC model the variations of vocal tract using linear and logarithmic filters. At low frequencies linear filters uses to capture the information present in spoken utterance and at high frequencies logarithmic filters use to capture the information. In MFCC below 1000Hz frequency spacing treated as linear and above 1000Hz as logarithmic. MFCC coefficients are generated using following steps.

1. Segment speech signal using 10ms window and perform Short term fourier transform(STFT) using 30-ms window.
2. Used triangular filter bank based on mel frequency scale to wraped the magnitude spectrum.
3. Based on each filter out log energy is computed.
4. Performed the Discrete cosine tranforms(DCT) on the filterout and finally we calculate cepstral coefiicients.

Residual cepstral coefficients(RCC) extracts from LP residual of speech signal. Higher frequencies of LP residual represent the excitation source information. RCC calculates using following steps.

1. Perform the fourier transfrom on the speech signal to transform the signal from time domain to frequency domain
2. Performs the logarithmic operation.
3. Perform inverse fourier tranform to calculate cepstral coefficients.

Higher frequencies to the cepstral coefiicients called as RCC.

Over the recent years an extensively research has been done on language identifation problem. Torss-Carrasquillo et al proposed an approach which based on gaussin mixture model. In this work they use shifted delta cepstral features in 7-1-3-7 manner. I-vector based approaches are the curent state of art for speech recognition and languagae identifiction task. An i-vector is the compact representation of speech utterance using universal background model. Typical feature length of i vector is 400-600. In the i vector based approaches main issue is to perform feature extraction and classification seprately as it computationally expense process.

Deep neural network(DNNs) based approaches are recently very popular is different applications such as visual object recognition and acoustic modeling etc. DNNs and its variation are the current state of art for machine learning algorithms. DNNs based approaches perform better then previosly GMM based approache. DNNs based approach remove the constraint regarding the data distribution which present the GMM. Hinton et al [1] first successfully use DNNs for acoustic modeling. First time Lopez-Moreno [2] show that DNNs performs better compare to state of art i-vector based systems. Richardson et al [3] trained single DNN to perform task of Speech recognition

978-1-5386-5457-6/18/$31.00 ©2018 IEEE

and LID simultaneously. They shows better results compare to previous approaches.

Nandi et al [4] uses GMM based classifier for LID task. In this work they propose to use complementry information presents in MFCC and RCC. In this work we performs decision level fusion approach for LID. Vocal tract information represented by MFCC features and excitation source information represented by RCC features. We use DNN based end to end approach for this task.

Rest of the paper organised as follows: In section 2 we explain DNN, In section 3 proposed approach, In section 5 results and experiments explains in details and section 6 discuss about conclusion.

ese components, incorporating the applicable criteria that follow.

II. Deep neural network(DNNs)

Now a days deep learning technique is very popular in the field of speech recogonotion. A regular DNN is a frame based classifier [5] where the decision from every frame is given equal inportance and the average of all the frames would represent the final decision to the utterance. In deep learning we are using concept of dense layer, convolutional layer, recurrent layer , activation function , loss function and back propagation based training. DNN based approaches are also popular in computer vision and NLP based applications.

In deep learning a node is the basic unit and represent a scalar value. These nodes are not interconnected. Layer is a set of nodes and represents a vector. An activation function is applied to each node. The value of node is computed by weighted sum of the input followed by an activation function. In Deep learning architecture mainly three layers presents called as input layer, output layer and intermediate layer(hidden layers, hidden layers are usually multiple layers). Multiple layer is important because when they are combined network learn complicated relationship between input and output. The depth of the network is depends on total no. of hidden layers presents in a network and width of the layer is number of nodes in that particular layer. In deep learning , all layers are directly learned from the data. Deep learning is also called end – to end learning.

For designing a deep neural network, nunber of layers and loss function is problem specific. Output of any DNN depend upon input and weights associates with each input. A loss function is defined on difference between predicted output and actual output of each layer. To achieve the desired output weights should be adjusted in such a way that loss function should be minimum. In this work we are using 4 layer DNN architecture 700R500R200R100R manner.

III. Proposed Method

In this work we proposed a fusion based approach for automatic language identification task. Our idea is to combine complementry information represents by MFCC(Vocal tract) and RCC(excitation source) using DNN framework. In this work we are using decision level fusion approach. We performs our experiments on 13 indian languages data sets. Our algorithm is as follows

- Segment long duration speech signals into 5 second duration signal.
- Extract MFCC and RCC features of each segment of speech signal.
- Trained DNN using MFCC and RCC features individually for end to end learning
- Performs score level fusion on MFCC and RCC DNNs output.
- Test the final trained network using test data.

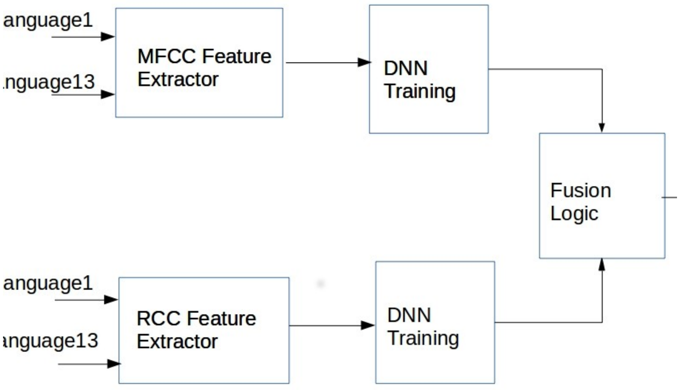


Fig. 1. Algorithm flowchart

Fig 1. shows the flow of our algorithm. More details about algorithm presents in next section.

IV. Experiments and results

A. Data Set

Our dataset consists of 13 Indian languages. The details of the structure of the data including the number of speakers per language in trained and tested data sets are given in Table 1. The three columns under the trained/tested data sets represent total speech data in number of hours, number of male and female speakers respectively. The style of speech data is read speech and has been recorded at 16 kHz sampling rate. For the purposes of our experiments, each sound file has been sliced into chunks of around 5s both in the training and testing datasets.

The datafiles are audios from various news Indian channels which have been sliced into three parts

- Training (most of the files)
- Testing
- Validation (around 60 from each language)

One unique challenge in building end-to-end LID systems on Indian language dataset is that most of the phonemes overlap amongst several languages. For instance, Telugu, Malayalam and Kannada being from the same language family have similar phonemes. The same can be said to be true for Assamese and Bengali. The geographical proximity also plays a role. Hence LR can be quite challenging on this dataset.

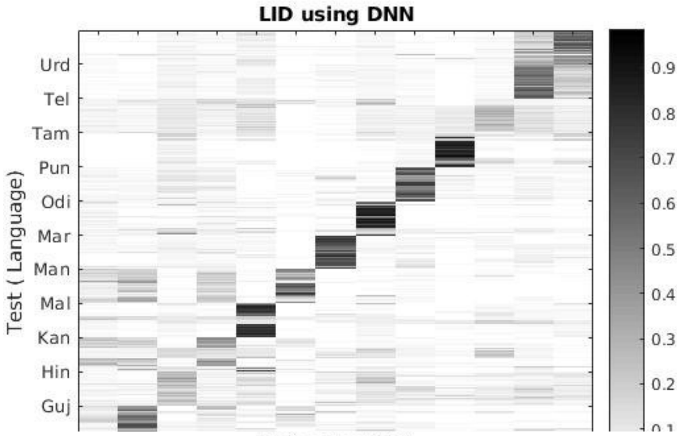
Language	**Train**			**Test**		
	Hours	***M***	***F***	***Hours***	***M***	***F***
Assamese	12:40	22	11	1.94	3	3
Bengali	9:91	24	35	1.53	15	15
Gujarati	9.71	175	15	2.18	37	36
Hindi	10.96	41	28	3.23	16	19

Language	Train			Test		
	Hours	*M*	*F*	*Hours*	*M*	*F*
Kanada	10.08	21	16	0.99	10	4
Malayalam	10.08	7	6	3.07	9	7
Manipuri	5.31	5	6	2.50	3	3
Marathi	7.84	74	31	2.47	17	15
Odiya	9.81	31	31	2.45	9	9
Punjabi	15.43	2	9	3.78	2	1
Tamil	10.43	21	21	3.15	4	4
Telugu	10.80	56	18	3.27	16	15
Urdu	11.00	10	9	4.10	5	4

TABLE 1. INDIAN LANGUAGE DATASET(M=MALE, F=FEMALE)

We extracted 39-dimensional MFCC and 40 dimensional RCC features from each of the 5 second chunks. DNNs are trained using a mini-batch stochastic gradient descent (SGD) with classical momentum. All the networks are trained to minimize the cross-entropy loss over the entire training set. For DNN, the frame based output were averaged before taking the final decision for the utterance.

In the first experiments we trained our DNN on MFCC and RCC features individually and compare our results with state of art i-vector based approach. MFCC based DNNs performs better than i-vector based approach. Fig 2. and Fig 3 shows the confusion matrix for RCC and MFCC respectively. Table 2 compare the MFCC with DNN results with state of art i-vector based approach and MFCC performs better compare to i-vector for almost every individual language and overall also. MFCC with DNN gives average ERR is 9.86% compare to 13.59% ERR for i-vector with 13 dimensions and 10.18% ERR with 39 dimensional vector.



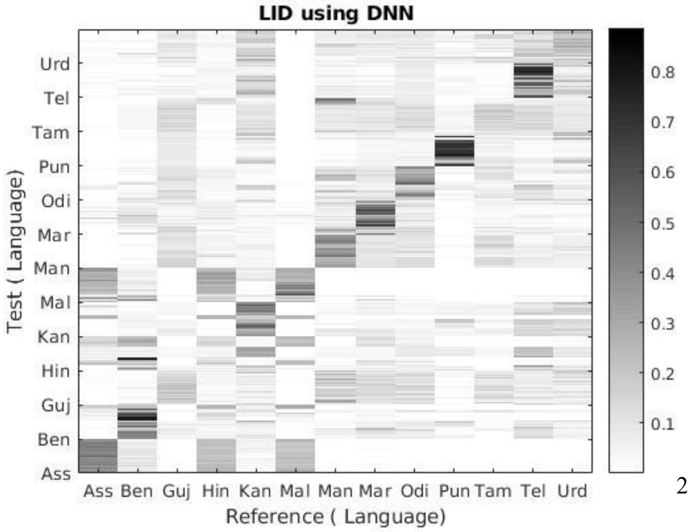


Fig. 2. Confusion matric for RCC

Fig. 3.

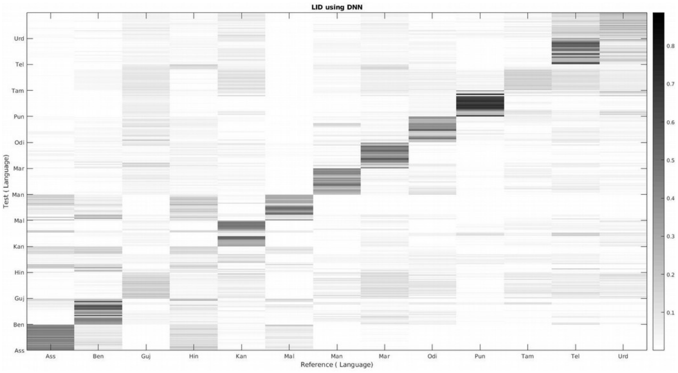


Fig. 4. Confusion matrix for MFCC

Confusion matrix for RCCConfusion matrix for min fusion

Language	I vector 13D	I vector 39D	DNN 13D	Our Approach
Assamese	6.05	5.12	3.78	4.80
Bengali	14.1	13	7.59	7.24
Gujarati	18.1	13.2	10.71	11.12
Hindi	26.1	19.9	25	21.26
Kanada	16.2	14.6	16	14.42
Malayalam	10.6	7.51	23.97	13.86
Manpuri	5.25	4.40	7.18	4.05
Marathi	13.8	10.1	15.40	8.20
Odiya	13	6.49	7.53	5.14
Punjabi	13.5	7.16	5.60	9.00
Tamil	26.9	19.9	27.78	19.40
Telugu	5.79	4.99	3.72	4.26
Urdu	6.94	5.73	8.06	5.44
Average	13.59	10.18	12.49	9.86

TABLE 2. ERR % of different methods where MFCC as feature

In the second experiments we perform the different fusion approaches such as Max fusion, Min Fusion, Average fusion and Min-Max fusion. We perform decision level fusion approach. After performing of MFCC and RCC individually ,we fuse DNN scores of these two features. The advantage of fusion is to use complementry information present in RCC and MFCC as both represent the different parts of speech production system. Table 3 shows the results of different fusion schems and Min-Max based performs better compare to all fusion schemes and also perform better than individually MFCC and RCC . MFCC and RCC individually gives ERR 9.86 % and 14.08 % respectively while Min. Max, average and min-max based fusion approaches gives ERR 10.98 %, 10.66 % , 10.39 % and 9.64 % respectively.

V. DISCUSSION AND CONCLUSION

In this work we represents vocal tract system using MFCC features and excitation source using RCC features. We combines these two complementry information using different fusion approaches. In this way we represent the full speech production system. Min-max based fusion approach gives

9.64% ERR which is better than individual features MFCC and RCC as their ERR is 9.86% and 14.08%.These results shows that Fusion approach performs better compare to individual features MFCC and RCC. As a part of future work, we try to do feature level fusions for MFCC and RCC, so that we trains only one DNN insted of two DNNs. We also wants to explore different DNNs architecture such as r-CNN and LSTM for LID.

Language	Min	Max	Average	Min-Max
Assamese	2.66	4.36	3.8	3.72
Bengali	6.69	8.40	7.05	6.12
Gujarati	11.4	14.78	12.28	11.10
Hindi	29.95	28.20	28.40	26.16
Kanada	12.8	10.27	12.09	13.12
Malayalam	12.08	12.27	11.60	10.20
Manipuri	5.22	4.14	4.18	4.15
Marathi	8.00	9.60	8.60	7.36
Odiya	7.36	5.37	5.12	5.41
Punjabi	11.70	11.20	11.40	10.36
Tamil	19.40	19.80	19.80	17.16
Telugu	5.73	4.09	4.42	4.18

Language	Min	Max	Average	Min-Max
Assamese	2.66	4.36	3.8	3.72
Urdu	9.75	5.98	6.42	5.87
Average	10.98	10.66	10.34	9.64

VI. Refrences

[1] Hinton, Geoffrey, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior et al. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups."*IEEE Signal Processing Magazine* 29, no. 6 (2012): 82-97.

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

[3] Richardson, F., Reynolds, D. and Dehak, N., 2015. A unified deep neural network for speaker and language recognition. *arXiv preprint arXiv:1504.00923*

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

[5] Choi, Keunwoo, György Fazekas, Kyunghyun Cho and Mark B. Sandler. “A Tutorial on Deep Learning for Music Information Retrieval.” *CoRR* abs/1709.04396 (2017):