The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages
29-31 August 2018, Gurugram, India

Analysis and Comparison of Features for Text-Independent Bengali Speaker Recognition

Shubhadeep Das
Department of Computer Science and Engineering
Indian Institute of Technology Guwahati
Email: shubhadeep.das@iitg.ac.in

Pradip K. Das
Department of Computer Science and Engineering
Indian Institute of Technology Guwahati
Email: pkdas@iitg.ac.in

***Abstract*—Speaker Recognition is the collective name of problems given to identifying a person or a set of persons using his/her voice. Variation of speaker speaking styles due to different languages can make speaker recognition a difficult task.**

In this paper, the main aim was to develop a system and compare different efficient text-independent Bengali speaker recognition systems that can give good rates of accuracy (greater than 90%) with not more than 10 minutes of speech data available for each speaker and can easily produce results without long amounts of delay. The experiments were carried out using the SHRUTI Bengali speech database and validated using TED-EX database.

We have also analyzed different features of a Bengali speaker using GMM-UBM framework, Joint Factor Analysis, i-vectors, CNN and RNN. Elaborate comparisons and classifications are carried out based on training durations and languages spoken by the speakers.

***Index Terms:* Text- Independent Speaker Recognition, GMM, UBM, LPCC, MFCC, ivectors, Bengali Speaker Recognition, CNN, RNN.**

I. INTRODUCTION

Identifying speakers using a specific language is an emerging area in speech processing. This is because of the hands-free scenario that all devices are gradually migrating to and only English speakers are well conceptualized. However, no model can give a 100% accuracy for a speaker across all languages. This is due to the ambiguous nature of speech itself. The way a speaker talks can vary significantly from time to time and depends on lots of factors like language spoken, emotion, environment and nature of listener. This makes speaker recognition a very challenging task. It is even more challenging for Indian regional languages due to the vast amount of diversity in the phonemes.

We have used a distribution of Gaussian functions to map each speaker features as suggested first by Reynolds [1]. Each such speaker has a different co-variance matrix and mean. This is essential as phonemes have different lengths in different languages. According to Auckenthaler [2], the GMM approach is more efficient than phoneme-based HMMs on text-independent speech. ALIZE Toolkit [3] which is used for the set of experiments conducted here, currently only uses GMM. Convolutional Neural Networks and Recurrent Neural Networks usually requires a large amount of data for predicting a speaker properly. For Indian regional languages no analysis for reducing the time delay using Neural Networks has been carried out. Thus substantial amount of study has to be done in Bengali speech recognition and an efficient framework for Bengali speaker identification is an emerging area of interest.

A. Toolkit

ALIZE toolkit is an open source platform for speaker recognition. Extraction of speech features like MFCC and i-vectors, training the model and testing of a new speech utterance, etc. are done by using this toolkit. For carrying out experiments using Neural Networks, Keras Framework of Python was used [4].

B. Database

For carrying out this experiment two databases have been used. For Bengali speech, data from IITG Kharagpur's SHRUTI speech corpus has been used [5]. The users were all natives of West Bengal and were from different backgrounds. The database is phonetically rich and can be categorized into different attributes. Below is the description of the database:

Items	Levels	Male	Female
Age/Gender	16-30	22	5
	31-40	4	3
Education	Undergraduate	2	0
	Graduate	22	8
Accent Category	D1	10	5
	D2	14	3

Fig. 1: Description of SHRUTI Database

For English speech, data from TED-ex have been used [6]. TED-ex databases contains speech of speakers from all over the world from the talk show TED Talks. 30 speakers has been chosen from this database such that their ages and genders are well distributed.

II. METHODOLOGIES USED FOR SPEAKER RECOGNITION

A. MFCC and LPCC Extraction

The first step is to extract voice characteristics from a speech signal. This is achieved by converting the given audio signal into vectors of features by a vector quantization technique. The whole signal is divided into small frames with the assumption that the characteristics of speech are invariant

10.21437/SLTU.2018-57

over that time period. Over each small frame a Hamming Window is applied. Then the auto-correlation coefficients are calculated for each small frame, ranging over a duration of 20 to 60 ms. The cepstral coefficients are obtained from the auto-correlation coefficients using Durbin's Algorithm. Using these cepstral values we derive Mel-Frequency Cepstral Coefficients (MFCC) and Linear Predictive Cepstral Coefficients(LPCC) [7].

The MFCC feature extraction process is shown in Figure 2.

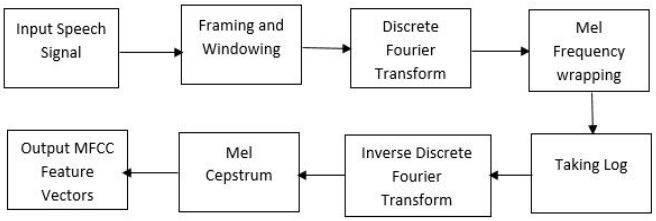

Fig. 2: Process of MFCC Extraction

B. Gaussian Mixture Models

GMMs are efficient in modeling multimodal distributions. Gaussian Mixture Model can correctly approximate a given waveform with sufficient number of mixture components which enables us to model the broad phonetic components of the voice of a speaker. The Gaussian Mixture Density Model consists of M component density tuples and each component has three parameters- mean vector, mixture weight and covariance matrix [8]. Gaussian Mixture Model of C component Gaussian densities is represented as follows:

$$p(x|\lambda) = \sum_{i=1}^{M} w_i g(x|\mu_i, \Sigma_i) \tag{1}$$

where, x is a vector that is obtained from the data per frame using MFCC.

$$w_i, i = 1, 2, 3, ...C$$

are the mixture weights and

$$g(x|\mu_i, \Sigma_i)$$

i = 1, 2, 3, ..., n are the component Gaussian densities.

C. Front End Factor Analysis (i-vector extraction)

The JFA modeling defines two separate spaces: One is the speaker space (defined by the eigenvoice matrix V) and the other is channel space (represented by the eigenchannel matrix U). But, it was observed that the speaker specific factors and channel factors are not completely independent [9]. Hence a new space named as total variability space was proposed. It contains speaker and channel variabilities simultaneously. Given an utterance, we can write the following:

For a particular utterance, we can write as follows:

$$M = m + Tw \tag{2}$$

where, M is a speaker and channel dependent supervector. m is a speaker and channel independent supervector or UBM supervector. T is a low rank matrix representing the principal directions of the speaker and channel variability and w is a vector representing the total factors. Components of w are called Total Factors (generally known as ivectors)[10].

D. Cosine Scoring

Cosine Scoring or Cosine Similarity is the most wildly used characteristic for high-dimensional positive spaces. The main reason for the popularity of the cosine similarity is that it can be evaluated very efficiently, mainly with sparse matrices or vectors. The cosine score is a parameter for measuring the similarity between two non-zero vectors which measures the cosine of the angle between the two vectors. Cosine scoring on ivectors w_1 and w_2 is obtained as follows [11]:

$$score(w_1, w_2) = \frac{w_1 * w_2}{|w_1||w_2|} \tag{3}$$

If two speech utterances are similar then their i-vectors will point in approximately same direction and angle between the vectors will be close to 0 which will give score near to 1. Two dissimilar speech utterances will give score near to -1 for the same reasons.

E. Spectrum feature description for Neural Networks

For the next set of experiments the input speech signal is broken into its component parts. The speech signal is broken out into the low-pitched parts, the next-lowest-pitched-parts, and so on. The number of divisions is again varied in experiments. Then by adding up how much energy is in each of those frequency bands (from low to high), we create a fingerprint of sorts for this audio snippet.

F. Recurrent Neural Networks (RNN)

A Recurrent Neural Network (RNN) is a class of artificial neural networks where connections between nodes form a directed graph along a sequence. It allows us to map the dynamic temporal behavior for a speech signal like the rate at which a speaker speaks. Unlike feed forward neural networks like CNN's, RNNs can use their internal state (memory) to process sequences of inputs. These type of Neural Networks are suitable for such context where a given part of input is related to input that comes before and after it [12].

G. Neural Network Description

A very simple variation of Neural Network has been used for carrying out these experiments. They are as follows:

- Features used: LPCC, MFCC and Spectral Feature
- Number of classes: 30, each corresponding to one speaker
- Batch Size: 100
- Number of Epochs: 50
- Number of Hidden Layers: 3(32 neurons, 48 neurons and 128 neurons)
- Activation functions used: We have tried with several functions like Sigmoid, RELU, SELU. Among these SELU activation function gave the best possible results
- Dropout Factor: 0.25
- Optimizer Used: AdaDelta Optimizer of Keras framework

III. Results

A. Training and Testing Description

- For training the UBM, alternative speech that resembles the target of the system was first collected and combined in a single file. For training in case of Neural Networks, the speech signal was converted into a 3-Dimensional format using a convolutional filter.
- For testing different methods were followed based on the features used. When ivectors was used as feature, cosine scoring technique is used to map a speaker whereas for MFCC and LPCC as a feature, the maximum likelihood probabilities was used as a metric. Testing was carried out for 10 seconds of random utterance for each speaker from both the databases. About 250 test cases were considered for each speaker. The worst case accuracy was taken into consideration for the purpose of accuracy calculation.

B. Using Gaussian Mixture Model as Speaker Model and LPCC as a Feature

The change in accuracy with varying training duration is shown in Figure 3 for Bengali speakers when LPCC is used as a feature with the GMM-UBM Framework.

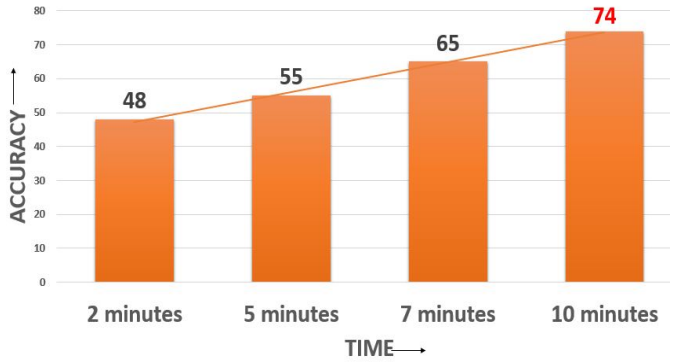

Fig. 3: Variation of accuracy % with training duration for LPCC and GMM for Bengali speakers

C. Using Gaussian Mixture Model as Speaker Model and MFCC as a Feature

Mel Frequency Cepstral Coefficients form a cepstral representation where the frequency bands are not linear but distributed according to the Mel-Scale. The change in accuracy with varying training duration is shown in Figure 4.

D. Using i-vector as a feature and cosine scoring as a classifier

I-vectors convey the speaker characteristic among other information such as transmission channel, acoustic environment or phonetic content of the speech segment [13]. The results are shown in Figure 5.

For the above set of experiments, English speakers showed a much higher average accuracy of 84%.

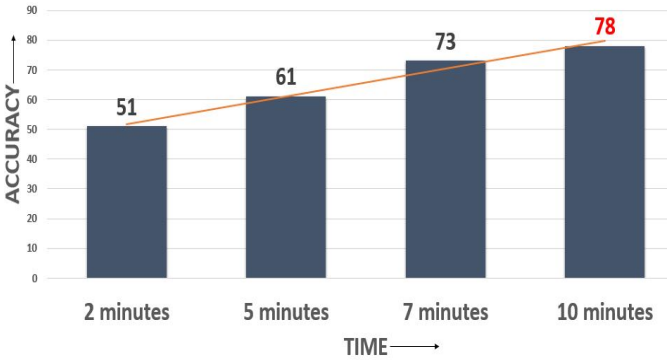

Fig. 4: Variation of accuracy % with training duration for MFCC and GMM for Bengali speakers

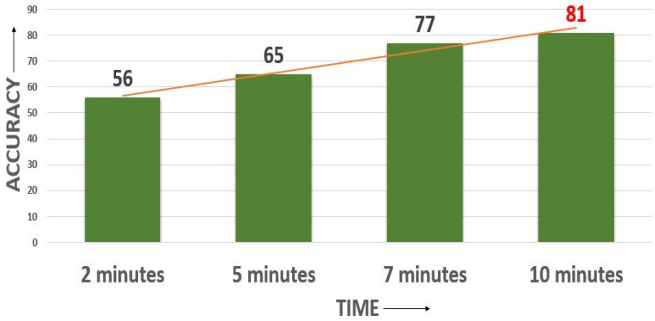

Fig. 5: Variation of accuracy % with training duration with ivector for Bengali speakers

E. Effect of varying size of Overlapping Windows

Speech is a non-stationary signal where properties change quite rapidly over time [14]. So for most phonemes, the properties of the speech remain invariant for a short period of time (5-100 ms) [15].

For these set of experiments, we have applied the Hamming Window with an overlap of 80 samples and frame size being fixed at 320 samples. To successfully identify a speaker we need to include as much information of the user as possible. Applying a window of any type does some type of approximations. In case of Hamming Window it is assumed that the signal outside the window has a value 0 [16]. Keeping this in mind we have changed the overlapping window size to 10, 20, 30 and so on to see the effect on the accuracy obtained.

The results described in Figure 6 and Figure 7, shows us that variation of phonemes for a speaker remains invariant for a much smaller amount of time in Bengali speech as compared to English language. Thus larger (more than 200) overlapping window size as well as frame size is not favourable to capture speaker characteristics for the languages with more variations than English. Smaller window size of less than 30 samples yields maximum accuracy for all the models in case of Bengali language.

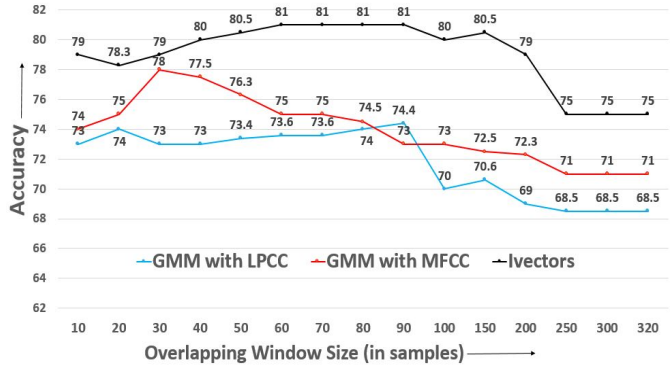

Fig. 6: Effect of accuracy % by changing overlapping window size on various models for Bengali Language

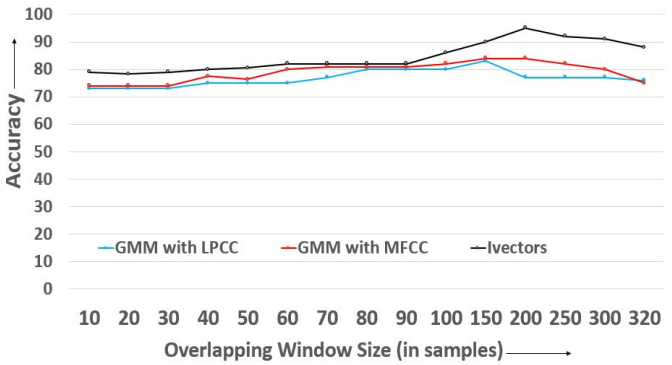

Fig. 7: Effect of accuracy % by changing overlapping window size on various models for English Language

F. Results obtained using Covolutional Neural Networks (CNN)

The best results obtained using Convolutional Neural Networks is described here speaker wise for Bengali language.

TABLE I: Accuracy for 30 Bengali speakers with 10 minutes of speech data and different features with varying Epochs in CNN.

Number of Epochs	Accuracy % with MFCC	Accuracy % with Spectrum of Energy
10	71	79
30	78	79
50	78	80.5
70	82.2	81
90	83	90
100	83	90
150	85	91.2
200	85	91
Average Accuracy	**72.3**	**86**

G. Results obtained using Recurrent Neural Networks (RNN)

The optimal results obtained using Recurrent Neural Networks is described here speaker wise. Similar to CNN, two features are used here: spectrum wise energy values in each band along with MFCC. Number of epochs is fixed to 50. The activation function used is SELU [17].

TABLE II: Accuracy for 30 Bengali speakers with 10 minutes of speech data and different features with varying Epochs in RNN

Number of Epochs	Accuracy % with MFCC	Accuracy % with Spectrum of Energy
10	79	90
30	86	92
50	87	95
70	88	96
90	88.2	96
100	89	96.5
150	89	97
200	89	97
Average Accuracy	**86**	**94.9**

This shows us that there is a significant improvement of accuracy for RNN as compared to CNN. The improvement also comes for some speakers whose accuracy has been low throughout for all the other models developed before.

IV. CONCLUSIONS

All the experiments for all the models were carried out in two parts, one using SHRUTI database for Bengali Language and one using TED-ex database for English languages.

- For GMM based systems, the accuracy of English speakers are high over Bengali speakers. For training of low duration English Language shows an accuracy of 81% over 48% of Bengali Language. This shows for low amount of data, Bengali speakers may be difficult to characterize.
- For ivectors based systems, English speakers have a high accuracy (96%) which is even better than Convolutional Neural Networks (89%) for 10 minutes of speech data. Bengali speakers on the other hand has a better (86%) accuracy for CNN as compared to ivectors (81%). The accuracy of English speakers improved (97%) for Neural Networks when RNN was introduced.
- When size of frames and size of overlap for clamping windows are increased, English speakers showed steady accuracy rate (deviation by 2%) for overlapping windows of small (less than 50 samples) size. The accuracy of some Bengali speakers increased by about 10% for windows of small size. This shows Bengali speech may be more sensitive to small changes and the length of the phonemes are small.

To conclude, English languages seems to be performing best in ivector based systems and Neural Network based systems provided enough speech data is available for each speaker. Bengali speakers work best with Neural Networks but the feature extraction has to be over small lengths of windows. However, the above results may be also due to some anomaly in the databases chosen and must be repeated with several other datasets to reliably conclude the above points.

REFERENCES

[1] D. A. Reynolds, Speaker identification and verification using gaussian mixture speaker models, Speech Communication, vol. 17, no. 1-2, pp. 91108, 1995.

[2] Auckenthaler, R., Carey, M., and Lloyd-Thomas, H., Score Normalization for TextIndependent Speaker Verification Systems, Digital Signal Processing 10 (2000), pp. 42-54.

[3] "ALIZE-Speaker-Recognition/alize-core", GitHub, 2018. [Online]. Available: https://github.com/ALIZE-Speaker-Recognition/alize-core. [Accessed: 22- Jun- 2018].

[4] "Keras Documentation", Keras.io, 2018. [Online]. Available: https://keras.io/. [Accessed: 22- Jun- 2018].

[5] B. Das, S. Mondal and K. Das, "Shruti Bengali Bangla ASR Speech Corpus", Cse.iitkgp.ac.in, 2018. [Online]. Available: http://cse.iitkgp.ac.in/ pabitra/shruti_corpus.html. [Accessed: 22- Jun- 2018].

[6] . Agi and N. Ljubei, "TED talks — Natural Language Processing group", Nlp.ffzg.hr, 2018. [Online]. Available: http://nlp.ffzg.hr/resources/corpora/ted-talks/. [Accessed: 22- Jun-2018].

[7] L. Rabiner and B. Juang, Fundamentals of speech recognition, 3rd ed. Englewood Cliffs, N.J.: PTR Prentice Hall, 1993, pp. 24-29.

[8] D. Reynolds, Gaussian Mixture Models, Encyclopedia of biometrics, pp. 827832,2015

[9]] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, Front-End Factor Analysis for Speaker Verification, Audio, Speech, and Language Processing, IEEE Transactions on, vol. 19, no. 4, pp. 788798, May 2011.

[10] P. Verma and P. K. Das, i-vectors in speech processing applications: a survey, International Journal of Speech Technology, vol. 18, no. 4, pp. 529546, 2015.

[11] N. Dehak, P. A. T. Carrasquillo, D. Reynolds, R. Dehak, Language Recognition via Ivectors and Dimensionality Reduction, pp. 857-860, INTERSPEECH 2011.

[12] Altosaar, T. and Meister, E. (1995) Speaker Recognition inEstonian Using Multi-Layer Feed-Forward Neural Nets. Pro-ceedings EUROSPEECH 1995, vol. 1, pp. 333-336.

[13] D. Reynolds, Universal Background Models, Encyclopedia of Biometrics, pp. 1547 1550, 2015.

[14] S. Ahmadi and A.S. Spanias, Cepstrum-Based Pitch Detection using a New Statistical V/UV Classification Algorithm, IEEE Trans. Speech Audio Processing, vol. 7 No. 3, pp. 333-338, 1999.

[15] Andrews, W., Kohler, M., Campbell, J., Godfrey, J., Hernandez-Cordero, J., 2002. Gender-dependent phonetic refraction for speaker recognition. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2002), Vol. 1, Orlando, Florida, USA, May 2002, pp. 149-152.

[16] Aparna R, Chithra PL. An effective method for continuous speech segmentation using filters. National Conference on Computing and Intelligence Systems. 2012; 1(1):pp. 1723.

[17] D. GUPTA, "Fundamentals of Deep Learning Introduction to Recurrent Neural Networks", Analytics Vidhya, 2018. [Online]. Available: https://www.analyticsvidhya.com/blog/2017/12/introduction-to-recurrent-neural-networks/. [Accessed: 22- Jun- 2018].