Gyancity Journal of Electronics and Computer Science,
Vol.5, No.1, pp. 12-18, March 2020
ISSN: 2446-2918 DOI: 10.21058/gjecs.2020.51002

Mechanism for Ensuring Teacher's Presence in Classroom Using Deep Learning

[1]RIZWANA MAHAR, AND [1]GHULAM MUSTAFA MEMON
[1]Department of Software Engineering, Mehran University of Engineering and Technology, Jamshoro, Pakistan.
rizwanamahar83@gmail.com, mustafa.09sw77.muet@gmail.com

Abstract

Education is one of the main components for the growth of a state, and there is not much research done in order to improve its quality especially in Pakistan. To improve Education quality the starting point is to boost Teaching Mechanism especially Government Schools where teacher are generally not present in class. Biometrics is not a valid system to use, as Teachers have found ways to tamper this type of attendance system. Using camera raises privacy issues as well as significant man or machine power is required to solve this problem. This paper presents a speaker recognition system applied to the problem of teacher identification in "ghost schools". The system uses a neural speaker embedding system that maps the audio lectures to a hyperspace where teacher similarity is measured by a cosine distance. We also present a corpus of audio lectures collected from 5 different teachers of 5 different courses. The dataset can be used for the evaluation of teacher identification system and contains audio lectures both for enrollment and testing. Our proposed system achieves an accuracy of 67% on the test set of the above mentioned corpus. We have made our code and corpus publicly available for reproducible research.

Key Words: Speaker Recognition, Speaker Verification, Teacher Identification.

I. INTRODUCTION

Speaker Recognition is the task of identifying person's identity from its voice prints. The voice prints of the individuals vary significantly due to difference in the vocal tract length, shape and other speech production organs. Furthermore, a part from physical difference, rate and manner of speech, intonation, accent and choice of vocabulary also very drastically among individuals. Hence, it is important for a speaker recognition system to take these features into account in order to do accurate recognition. Speaker Recognition can operate in two modes i.e. verification and identification. A verification system determines whether or not a person is who they say they are (i.e. the person claims an identity and the system tries to prove whether or not that claim is true). On the other hand, an identification system attempts to

ISSN: 2446-2918 DOI: 10.21058/gjecs.2020.51002

establish a person's identity from scratch (i.e. the system tries to associate a person with an identity from a set of identities in the system's database). Both speaker verification and identification can be text-dependent (gnostic to what is said) as well as text-independent (agnostic to what is said). One of the important application of speaker recognition system lies in forensics or biometrics. Researcher are trying to integrate speaker recognition technology in criminology to supplement auditory analysis (Niemi-Laitinen et. al. [1], Thiruvaran et. al. [2]). Education is another area where speaker recognition technology can benefit and this is what the main crux of this research work. We present a speaker recognition system for education sector to automatically identify teacher's identity from the audio lectures and pin point "ghost schools". We also present an evaluation corpus of audio lectures containing lecture recordings from 5 different teachers of 5 different subjects. The rest of the paper is structured as follows: Section 2 discusses the related work. Section 3 describes main components of a speaker recognition system. Section 4 discusses our proposed system. Section 5 and 6 are experimental setup and conclusion respectively.

II. RELATED WORK

Previously, identity vector or i-vector based approaches have been used extensively to model inter-speaker variability. These approaches use simple distance metric such as cosine distance or more sophisticated techniques such as PLDA (Probabilistic Linear Discriminant Analysis (Prince et. al. [3], Matějka et. al. [4] Cumani et. al. [5]) to perform classification.

DNNs (Deep Neural Networks) have also been used to replace some components of traditional automatic speaker recognition. One such approach is to first extract bottleneck features from a DNN and the train GMM on these bottleneck features to extract i-vectors (Ghalehjegh et. al. [6]). In another work authors have utilized speech recognition neural network to extract posteriors for i-vectors instead of a GMM-UBM (Gaussian Mixture Model-Universal Background Model) (Lei et. al. [7]). Variani et. al. [8] used the final layer activations of a trained frame level DNN to create a speaker representation. They named their representation d-vector. These DNN based approaches have shown improvement over traditional i-vector based approaches.

Recently, there have been attempts to do E2E (End-to-End) automatic speaker recognition. LSTM (Long Short Term Memory) trained in E2E fashion achieved an accuracy of 98% on "Ok Google" benchmark for text-dependent speaker recognition (Heigold et. al. [9]). In another work authors have used same and different speaker pairs to train an E2E text-independent speaker recognition (Snyder et. al. [10] Li et. al. [11]). The achieved 13% relative improvement over i-vector based baseline.

Our approach is similar to that of (Li et. al. [11]) in a sense that it also uses same and different speaker pairs to learn speaker embeddings. However, we used a much simpler deep learning architecture.

Gyancity Journal of Electronics and Computer Science,
Vol.5, No.1, pp. 12-18, March 2020
ISSN: 2446-2918 DOI: 10.21058/gjecs.2020.51002

III. EXPERIMENTAL SETUP

This section discusses our experimental setup, dataset and results.

Speaker Recognition System

There are two main processes involved in a speaker recognition system: (1) enrollment and (2) testing/recognition. In the former speaker model is trained using the features extracted from the target speaker utterances while in the latter features are first extracted from an unknown utterance and then compared against speaker model(s) in the database to compute similarity. Finally, a decision is made based on the similarity score. A pipeline of a standard automatic speaker recognition system is shown in Fig. 1(a-b). The upper half illustrates the enrollment while the lower one depicts testing/recognition.

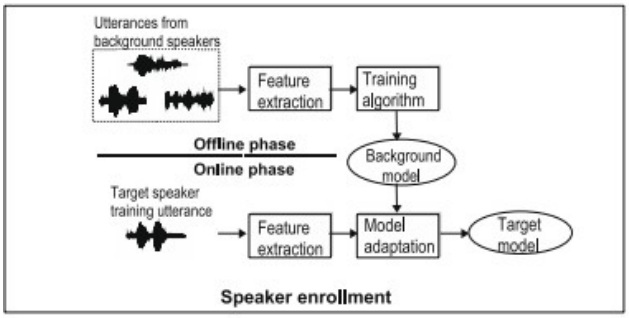

FIG. 1(a). COMPONENTS OF A STANDARD AUTOMATIC SPEAKER RECOGNITION SYSTEM. IN THE ENROLLMENT PHASE, A SPEAKER MODEL IS CREATED USING ONLY FEW TARGET SPEAKER UTTERANCES

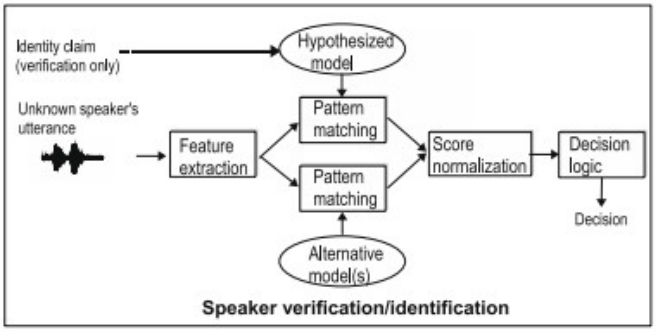

FIG. 1(b). IN THE RECOGNITION PHASE, UNKNOWN UTTERANCE IS COMPARED WITH THE SPEAKER MODEL(S) AND DECISION IS MADE BASED ON SIMILARITY SCORE

In the feature extraction process, speaker specific features are extracted from the raw speech signals. Some of the commonly used features are MFCCs (Mel-Frequency Cepstral Coefficients), Log-filter bank energies and PLP (Perceptual Linear Predictive) coefficient.

ISSN: 2446-2918 DOI: 10.21058/gjecs.2020.51002

Deep Speaker

We employ a simple deep neural architecture to do automatic speaker recognition. Our pipeline starts by first computing MFCCs from the raw audio signal. We compute 39-dimentional MFCCs over a window size of 25 milliseconds with a 10 millisecond overlap between the consecutive frames. The processed audios are then passed to feed-forward deep neural network for feature extraction. The 200 dimensional speaker embeddings are obtained by applying normalization to the extracted features. Finally, a triplet loss (Schroff et. al. [12]) is applied on speaker embeddings that minimizes the cosine distance between the speaker embeddings of the same speaker and maximizes the cosine distance between the speaker embeddings of different speakers. In order to avoid local minima we also employ Softmax pre-training. The detailed architecture is presented in Table 1.

TABLE 1. DEEP NEURAL ARCHITECTURE FOR AUTOMATIC SPEAKER RECOGNITION

Layer Name (Type)	Dimension	Parameter
Input (Input Layer)	390	00
Fully Connected (Dense)	200	78200
Normalization (Lambda)	200	00
Embeddings (Lambda)	200	00
Softmax (Dense)	105	21105

The cosine distance between two embeddings y_i and y_j is given by Equation (1):

$$\text{Cosine-Distance}\ (y_i, y_j) = 1 - y_i^T y_j \quad (1)$$

The triplet loss takes three embeddings: (1) an anchor i.e. embedding of particular speaker (2) a positive sample i.e. embedding of the same speaker (3) a negative sample i.e. embedding of the different speaker and updates in such a way that cosine distance between anchor and positive sample minimizes while cosine distance between anchor and negative sample maximizes. More formally,

$$s_i^{ap} - \alpha < s_i^{an} \quad (2)$$

Where s_i^{ap} is the cosine distance between anchor a and positive sample p, s_i^{an} is the cosine distance between anchor a and negative sample n and α is the minimum margin constant. The learning paradigm of a triplet loss is shown in Fig. 2.

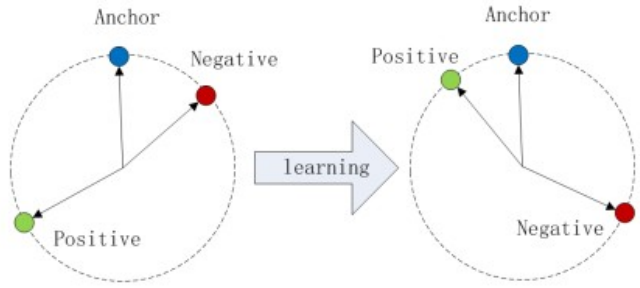

FIG. 2. TRIPLET LOSS LEARNING USING COSINE DISTANCE

Dataset

We have used VCTK (Voice Cloning Toolkit) corpus to train our background feature (speaker embeddings) extractor. The corpus consists of speech recordings from 105 English speakers where each speaker uttered around 400 sentences. For enrollment and testing we have collected our ₌speech data consisting of audio lectures from 5 different teachers of 5 different subjects. The details of this corpus are presented in Table 2.

TABLE 2. STATISTICS OF ENROLL AND TEST SET

Teacher	Subject	Enroll Utterances	Test Utterances
Barbara	General Knowledge	10	09
Rosy	Mathematics		15
John	Drawing		03
Daisy	English		08
Gabrielle	Science		05

Results

We have achieved an overall accuracy of 67.44%. Detailed class wise results for each speaker class are given in Table 3.

TABLE 3. CLASS WISE RESULTS FOR EACH SPEAKER

Teacher	Precision	Recall	F1
Barbara	0.89	0.89	0.89
Rosy	0.64	0.47	0.54
John	0.00	0.00	0.00
Daisy	0.64	0.88	0.74
Gabrielle	0.70	1.00	0.82

IV. CONCLUSION

This paper presents an application of automatic speaker recognition in education sector. Our proposed system automatically recognizes the teacher's identity from its voice print in audio lecture. We have also presented a speech corpus of audio lectures

ISSN: 2446-2918 DOI: 10.21058/gjecs.2020.51002

from 5 different speakers of 5 different subjects. Our code and corpus is available publicly for reproducible research.

The proposed system will help in maintaining teacher's attendance in schools through their voice identification. Implementing this mechanism in schools the education system can be improved greatly as teacher's play significant role in education sector.

V. FUTURE WORK

Further research can be done by detecting subject of teaching in voice print. Our work is limited for primary school level subjects and language dependent, hence working on these points can make this research more valuable for education area.

VI. ACKNOWLEDGEMENT

Praise to **ALLAH**, the most gracious and the most merciful. Without HIS blessings our accomplishment would have never been possible. Authors are also acknowledging to Institute of Information & Communication Technologies, Mehran University of Engineering & Technology, Jamshoro, Pakistan, for providing facilities to conduct this research paper.

VII. REFERENCES

[1] Niemi-Laitinen, T., Saastamoinen, J., Kinnunen, T., and Fränti, P., "Applying MFCC-Based Automatic Speaker Recognition to GSM and Forensic Data", Proceedings of 2nd Baltic Conference on Human Language Technologies, Tallinn, Estonia, pp. 317-322, April, 2005.

[2] Thiruvaran, T., Ambikairajah, E., and Epps, J., "FM Features for Automatic Forensic Speaker Recognition", Proceedings of 9th Annual Conference of the International Speech Communication Association, 2008.

[3] Prince, S.J., and Elder, J.H., "Probabilistic Linear Discriminant Analysis for Inferences About Identity", Proceedings of IEEE 11th International Conference on Computer Vision, pp. 1-8, October, 2007.

[4] Matějka, P., Glembek, O., Castaldo, F., Alam, M.J., Plchot, O., Kenny, P., and Černocky, J., "Full-Covariance UBM and Heavy-Tailed PLDA in i-Vector Speaker Verification", Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4828-4831, May, 2011.

[5] Cumani, S., Plchot, O., and Laface, P., "Probabilistic Linear Discriminant Analysis of i-Vector Posterior Distributions", Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7644-7648, May, 2013.

[6] Ghalehjegh, S.H., and Rose, R.C., "Deep Bottleneck Features for i-Vector Based Text-Independent Speaker Verification", Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 555-560, December, 2015.

[7] Lei, Y., Scheffer, N., Ferrer, L., and McLaren, M., "A Novel Scheme for Speaker Recognition Using a Phonetically-Aware Deep Neural Network", Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1695-1699, May, 2014.

[8] Variani, E., Lei, X., McDermott, E., Moreno, I.L., and Gonzalez-Dominguez, J., "Deep Neural Networks for Small Footprint Text-Dependent Speaker Verification", Proceedings of

IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4052-4056, May, 2014.

[9] Heigold, G., Moreno, I., Bengio, S., and Shazeer, N., "End-to-End Text-Dependent Speaker Verification", Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5115-5119, March, 2016.

[10] Snyder, D., Ghahremani, P., Povey, D., Garcia-Romero, D., Carmiel, Y., and Khudanpur, S., "Deep Neural Network-Based Speaker Embeddings for End-to-End Speaker Verification", Proceedings of IEEE Spoken Language Technology Workshop, pp. 165-170, December, 2016.

[10] Li, C., Ma, X., Jiang, B., Li, X., Zhang, X., Liu, X., and Zhu, Z., "Deep Speaker: An End-to-End Neural Speaker Embedding System", arXiv Preprint arXiv, [ISN: 1705.02304], 2017.

[11] Schroff, F., Kalenichenko, D., and Philbin, J., "Facenet: A Unified Embedding for Face Recognition and Clustering", Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 815-823, 2015.