Intelligent Speaker Identification System under Multi-Variability Speech Conditions

Banala Saritha[1][0000−0002−6499−3561], Tungala Thiru Venkata Naga Manoj[2], Sachin Kumar Sharma[3], Rabul Hussain Laskar[4], Madhuchhanda Choudhury[5], and Anish Monsley K[6]

National Institute of Technology Silchar, Assam 788010, India
banalasarita@gmail.com,tungala20_ug@ece.nits.ac.in, sachin20_ug@ece.nits.ac.in,rhlaskar@ece.nits.ac.in,m_chhanda@ece.nits.ac.in, anishmonsley@yahoo.com

Abstract. Speech is the natural source of information for human identification in most biometrics, forensics, and access control systems. Mismatch in Speech data is one of the biggest challenges preventing speaker identification systems from being employed in real-world scenarios. This research explores how intelligently speaker identification tasks are affected by degraded speech. Our preliminary investigation into mismatch effects in conversational style telephonic speech conditions utilizing the IIT-G database includes mismatch in sensor, environment,language and conversational style. Convolutional neural networks (CNNs) have surpassed traditional techniques in speaker identification (SI) systems in recent years. This paper proposes a novel architecture based on a VGG-like network for an end-to-end speaker identification system. The proposed architecture outperforms the statistical methods with an improvement of 7% accuracy in identifying the speakers. The results show that the suggested approach is more accurate than state-of-the-art speaker identification techniques and notable performance deterioration compared to the matched scenario.

Keywords: Speaker Identification, Deep Convolutional Neural Networks, Conversational Speech, Degraded conditions.

1 Introduction

Speech is a robust mode of communication that transmits rich and essential information such as gender, accent, emotion, and other distinctive traits of a speaker. These unique features help identify the speaker and find the applications across biometrics, forensics, and access control systems [1]. Finding an unknown speaker from a group of well-known speakers involves extensive research in speaker identification [2]. Researchers' attention has recently been to recent developments in deep learning techniques [3]. Speech is more unreliable than other forms of personal identification like the iris, face, and fingerprint. Age, gender, emotion, language, and speaking style concerns can affect speech

characteristics. The speech samples used for training and testing are one of the main variables limiting speaker identification accuracy in addition to these inconsistencies. The downsides of using human speech for identification include variability within a speaker and mismatch in testing and training situations [4]. A speaker's voice can change over time and may not always sound the same, referred to as intra-speaker variability. A mismatch occurs when speech acquires from the same person in diverse contexts. The extensive usage of several sensing devices (such as a microphone on mobile phones) and contextual factors may change the voice quality of the original speech. Also, multi-language and conversational styles affect the performance of speaker identification. To work around these shortcomings, we used the Multi-Variability (IITG-MV) speaker identification database [5]. We also introduced a novel structure for an end-to-end intelligent speaker identification system relying on a VGG-like network. The rest of the paper is organized as follows. The related works are described in Section II. The analysis and comparison of the experimental work with the acquired results are presented in Section IV.

2 Related Works

Due to advances in computing power and the emergence of massive databases, deep learning techniques for speaker identification have garnered much attention from researchers. Acoustic models based on Deep Neural Networks (DNN) have performed much better than those based on Gaussian Mixture Models over the past few years (GMM). The authors of [6] compared speaker identification algorithms that use DTW, GMM, and SVM. The authors [7] used a variety of normalizing strategies to overcome the discrepancy between the training and testing settings. The researchers employed feature mapping approaches in [8]. The authors introduce speaker model synthesis [9] in mismatched conditions. Additionally, factor analysis by authors [10] and nuisance attribute projection (NAP) [11] approaches are proposed. In their research, the effectiveness of short-time cepstral characteristics generated by filter banks with Bark and ERB rate warping is examined and the robustness of speaker identification in mismatch situations. Mahesh et al. [12] used Warped Filter Bank Features to accomplish the speaker identification task in mismatch contexts. They used databases with mismatched sensors for this. In [13], Haris et al. created the IITG Multivariability Speaker Recognition Database in the context of India. These authors used adapted Gaussian mixture modeling to assess the speaker identification and verification tasks. We also compared the findings of this research using the IITG-MV database. Moreover, we presented a deep CNN architecture for an end-to-end speaker identification system built on a VGG-like network.

3 Methodology

This study investigates the speaker identification task in an uncontrolled setting using input speech collected from various sensor devices. Phase III of the IITG

multi-variability speaker recognition database uses to evaluate our work. The Indian Institute of Technology Guwahati prepared the IITG-MV speech database. In the context of India, where speech is highly varied and diverse, the database promotes the reliability of speaker identification. Here, the third phase of the database is described briefly.

3.1 Multi-Variability Speech Data Base

IITG-MV Phase-III database: In the third recording phase, a telephone network was utilized, keeping in mind the possibility of remote person authentication in speech mode. At the same time, the Phase I dataset is used in the sensor mismatch scenario. Phase II database includes a multi-environment like laboratories, student rooms, corridors, etc. In contrast to phases I and II, a facilitator connects a conference call between two people in phase III during their leisure time. The Phase-III database contains four of the variabilities that are described below.

Multi-environment: Conversations were captured in various real-world settings, including coffee shops, offices, living spaces, laboratories, etc.

Multi-sensor: A sample frequency of 8 kHz is used to record speech data from several mobile phones.

Multi-lingual: Every speaker used their mother tongue or English (favorite language).

Conversation style: Over a conference call, each speaker spoke conversationally.

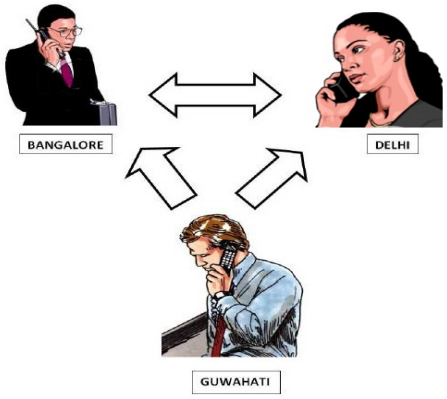

Fig. 1. Phase III data collecting scenario for a conference call [13]

The recording process is depicted in Figure 1. The individuals were asked to provide a meeting time or a specified time slot during which they and their friends would be available to speak. The facilitator then used his mobile handset

to call the individual at the appointed time. In response, the individual provided the contact number of the person he wanted to speak with. After putting the individual on hold, the facilitator called the other party. The facilitator then connected the two people on a conference call. There were no restrictions on the location or language of the recording. The entire time, subjects were discussed openly, with frequent language and subject matter shifts.

3.2 Proposed Architecture

This section presents a novel architecture for end-to-end intelligent speaker identification tasks built on a network similar to a VGG. The exceptional performance of the VGG architecture in speaker identification and computer vision is the inspiration for the proposed framework. In this study, we modified the layers and filters to redesign the fundamental structure of VGG-13. Fig. 1 depicts the proposed architecture. There are ten convolutional layers in this model. Each layer utilizes a rectified linear unit (ReLU) activation function and is a 2d convolutional layer with a kernel size of 3 x 3. After each convolutional layer, a batch normalizing layer is added on top. At the same time, the average pooling layer of stride 2 X 2 is added after every few convolutional layers. After a global average pooling (GAP) layer, a fully connected layer with 4096 neurons is implanted. A dropout of 0.5 is positioned on either side of the FC layer to prevent generalization issues. The softmax classification layer is eventually connected to the output convolutional layer. The 100 nodes of the output convolutional layer represent the number of speakers that can be identified using the softmax loss function.

4 Experimental Setup and Results

An intelligent end-to-end speaker identification is the main objective of this research. 20–30 msec of conversational style 1 speech data are used for training, while 8–15 msec of conversational style 2 are used for testing. Log-mel spectrogram is generated from each raw audio sample. The Librosa library is used in Python to carry out this operation for all voice recordings.

These spectrograms are scaled to 224 X 224 before being fed as input into the proposed architecture. Additionally, the presented model is trained on the Kaggle GPU using the TensorFlow Python library with a 256-batch size. The Adam optimizer was used for training, with the following learning rates: $lr = 0.001$, $\alpha = 0.95$, $\epsilon = 10^{-5}$. The speaker identification accuracy of the proposed model is measured as a percentage, which is the ratio of correctly predicted classes of test samples to the total positive prediction classes of the test samples.

Mismatch effects in conversational style:

The four variabilities of the Phase-III database are described in section 3.1. This work considers the mismatch effects in truly conversational-style in telephonic speech data during conference calls. For this purpose, single-sensor devices like mobile handsets, multi-environment conditions, and conversational

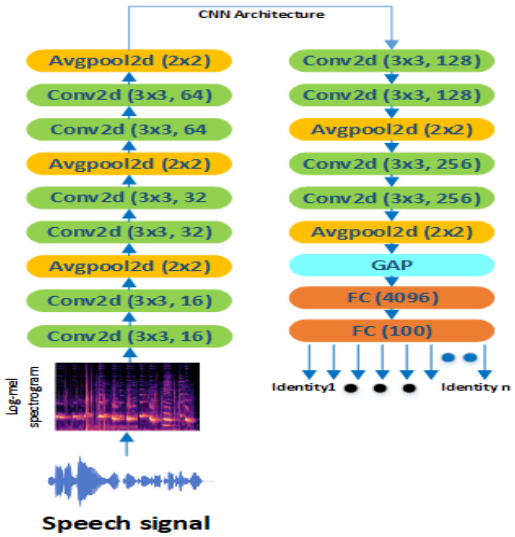

Fig. 2. Architecture of the proposed end-to-end intelligent speaker identification system

style with the mother tongue languages for two sessions are maintained throughout this work.

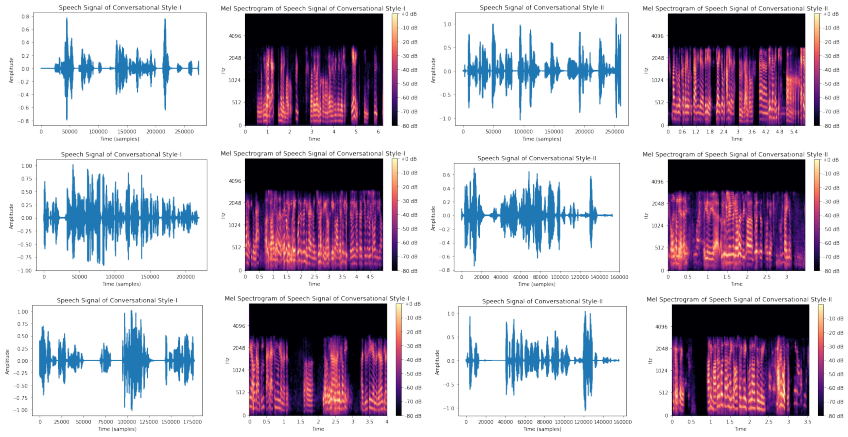

Fig. 3. Time doamin and Melspectrogram representation of speech data with different conversational style

This research considers conversational styles 1 and 2 speech data for training and testing purposes. Speaker identification accuracy was 98.05 percent and

95.62 percent for matched and mismatched styles, respectively. Speaker identification accuracy of 88.1% is achieved using traditional approaches [13]. Figure 4 displays the speaker identification accuracy of the proposed architecture under conversational style mismatch conditions.

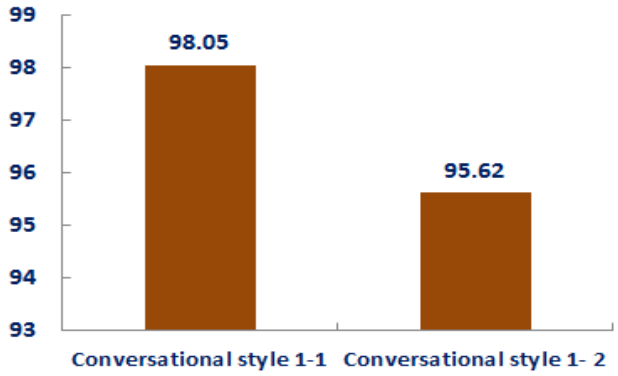

Fig. 4. Accuracy (%) comparison for conversational mismatch cases

5 Conclusion

The robustness of speaker identification systems in various circumstances and diverse situations is challenging for real-world applications. This research examined how speaker identification performance is affected by degraded speech. This analysis considers mismatch effects in telephonic speech data collected during casual conference conversations. We utilized Phase III of the IIT-G database to address the issues. The presented VGG-like architecture provides robust speech features for an end-to-end intelligent speaker identification system, in contrast to existing methods. It is evident from the results that the identification accuracy for conversational style mismatch in conference call scenarios has enhanced by 7%. Future research includes ResNet architectures to assess the speaker identification system performed when faced with language mismatch.

Acknowledgment

The authors gratefully acknowledge IIT Guwahati for providing the speech database. The authors also acknowledge the National Institute of Technology Silchar's Electronics and Communication Engineering Department for providing the resources required to complete this work.

References

1. B. Saritha, N. Shome, R. H. Laskar and M. Choudhury, "Enhancement in Speaker Recognition using SincNet through Optimal Window and Frame Shift," 2022 2nd International Conference on Intelligent Technologies (CONIT), 2022, pp. 1-6.
2. Saritha, B., Laskar, M.A., Laskar, R.H. (2023). A Comprehensive Review on Speaker Recognition. In: Biswas, A., Wennekes, E., Wieczorkowska, A., Laskar, R.H. (eds) Advances in Speech and Music Technology. Signals and Communication Technology. Springer, Cham. https://doi.org/10.1007/978-3-031-18444-4_1.
3. B. Saritha, M. A. Laskar, R. H. Laskar and M. Choudhury, "Raw Waveform Based Speaker Identification Using Deep Neural Networks," 2022 IEEE Silchar Subsection Conference (SILCON), Silchar, India, 2022, pp. 1-4, doi: 10.1109/SILCON55242.2022.10028890.
4. R. Jahangir, Y. W. Teh, H. F. Nweke, G. Mujtaba, M. A. Al-Garadi, and I. Ali, "Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges," Expert Syst. Appl., vol. 171, p. 114591, Jun. 2021.
5. Haris B C, Pradhan, G., Misra, A., Shukla, S., Sinha, R., Prasanna, S. R. M., 2011. Multi-variability speech database for robust speaker recognition. In: Proceedings of National Conference on Communications. pp. 1–5.
6. Jr Ding, Chih-Ta Yen, and Da-Cheng Ou, "A method to integrate GMM, SVM and DTW for speaker recognition," International Journal of Engineering and Technology Innovation, vol. 4, no. 1, pp. 38–47, 2014.
7. R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," Digital Signal Processing, vol. 10, no. 1-3, pp. 42–54, 2000.
8. D. Reynolds, "Channel robust speaker verification via feature mapping," in Proc. ICASSP '03, vol. 2, 2003, pp. II–53–6 vol.2.
9. R. Teunen, B. Shahshahani, and L. Heck, "A model-based transformational approach to robust speaker recognition," in Proc. Int. Conf. on Spoken Language Processing (ICSLP '00), vol. 2, 2000, pp. 495–498.
10. S.-C. Yin, R. Rose, and P. Kenny, "A joint factor analysis approach to progressive model adaptation in text-independent speaker verification," IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 7, pp. 1999–2010, 2007.
11. A. Solomonoff, W. Campbell, and I. Boardman, "Advances In Channel Compensation For SVM Speaker Recognition," in Proc. ICASSP '05,18 2005.
12. Chougule, Sharada Vikram and Mahesh S. Chavan. "Speaker Recognition in Mismatch Conditions: A Feature Level Approach." International Journal of Image, Graphics and Signal Processing 9 (2017): 37-43.
13. Haris B C, Pradhan, G., Misra, A. et al. Multivariability speaker recognition database in Indian scenario. Int J Speech Technol 15, 441–453 (2012).