Journal of Integrated
SCIENCE & TECHNOLOGY

Application of autoencoders in speaker recognition system in noisy environment

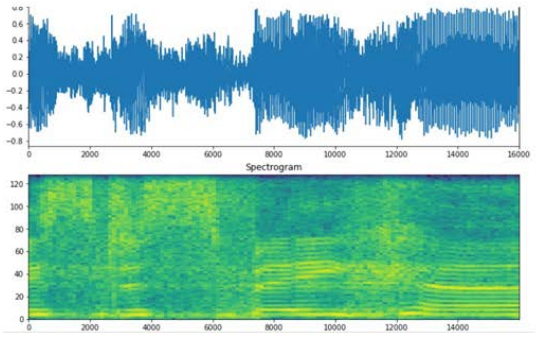
Arundhati Niwatkar,[1*] Yuvraj Kanse[2]

[1]*Shivaji University, Kolhapur, India.* [2]*KBP College of Engineering, Satara, Maharashtra, India.*

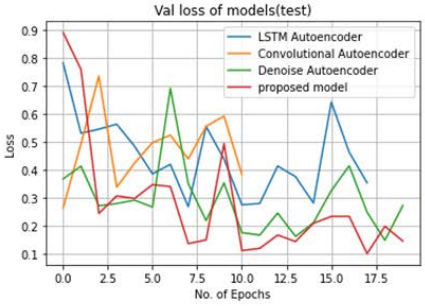
Received on: 27-Jul-2023, Accepted and Published on: 20-Oct-2023

Article

ABSTRACT



Spectrogram of Voice sample



Comparative loss rate of Autoencoders

One of the most difficult problems for autonomous speakers is speaker detection and identification because it requires clever technology for the creation of cutting-edge functioning systems. Traditional approaches for Speaker identification and recognition are inaccurate, time-consuming, and have low success rates. This work has been carried out to improve Speaker recognition and identification system accuracy while also increasing success rate. In this study, a dataset gathered from 5 speakers, both men and women, have been rationalized for evaluation of data. The gathered data have been utilized using the Data Augmentation approach based on the size of the dataset. The study implemented a system for recognizing and identifying speakers that makes use of deep learning and autoencoders, specially, in noisy environment. Additionally, in order to validate our result and to prove the novelty, the study compared the results with existing speaker recognition systems.

Keywords: Speaker recognition, Augmentation, Autoencoder, LSTM

INTRODUCTION

In recent years, notable advancements have been made in the field of Speaker Identification and Recognition, showcasing a growing depth of understanding among researchers. Leveraging individuals' vocal characteristics, this system enables the recognition of speakers. The realm of speaker recognition can be categorized into text-dependent and text-independent approaches. The process of identifying the trained sample speech that most closely corresponds to a speaker's voice is termed as speaker recognition.[1] Moreover, speaker recognition serves as a means of verifying or disproving the asserted identity of a speaker.

Within the landscape of traditional systems, methods such as Gaussian Mixture Models (GMMs), I-vectors, and Hidden Markov Models (HMMs) have played crucial roles in speaker recognition.[2] GMMs, for instance, represent the collective weights of Gaussian mixtures, constituting a probabilistic model. These models have exhibited notable success and accuracy in speaker identification. Meanwhile, contemporary trends point towards the integration of deep learning techniques in speaker recognition systems. Convolutional Neural Networks (CNNs) are currently the prevailing choice, though Autoencoders present a promising avenue for further advancements.[3] Autoencoder variants include Vanilla Autoencoders, De-noising Autoencoders, Variational Autoencoders, and more. In this study, a relatively limited dataset was utilized, sourced from five distinct speakers. Notably, the dataset was retained in its original, unprocessed form. Each utterance possesses a duration of 3 seconds. It is noteworthy that both training and testing were conducted with diverse texts, as the system employed is text-independent in nature.[4,5] A significant feature of this approach is its language-agnostic nature. Irrespective of the language of the training and testing data, the focus remains firmly on the inherent speech qualities of the speaker. The detailed expounded analysis of the designed methodology with details of experimental design, the outcomes of each experiment, and potential future applications of the designed model would prove beneficial in potential developmental applicability of current study.

*Corresponding Author: Arundhati Niwatkar
Email: amehendale@umit.sndt.ac.in

Cite as: *J. Integr. Sci. Technol., 2024, 12(2), 742.*
URN:NBN:sciencein.jist.2024.v12.742

 ISSN: 2321-4635
http://pubs.thesciencein.org/jist

Architecture Model

Figure 1 illustrates the designed model architecture for our research study. Our dataset consists of raw recordings obtained from 5 speakers, which is relatively small in size. To address this limitation, we have implemented data augmentation techniques, effectively expanding the dataset to provide a substantial amount of training data. This augmented dataset has undergone necessary preparations for further experimental analysis.

It's important to note that the speech samples collected from the speakers have undergone no preprocessing at this stage. After the initial database enhancement, the subsequent step involves transforming these voice samples into spectrograms.[6] To enhance the quality of these spectrograms, a morphological filtering process has been applied. Morphological filters, widely used in image processing, are adept at tasks such as image cleansing, enhancement, and preprocessing.[7] Specifically tailored for binary or grayscale images, these filters analyze object shapes and structures within an image. Their application helps eliminate noise, refine object outlines, and simplify or enhance image features.

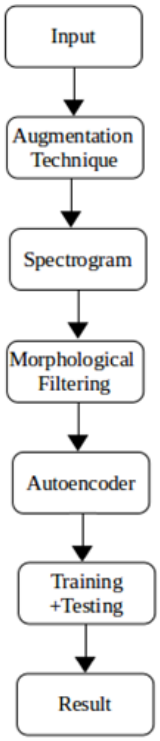


Figure 1. Proposed model for Speaker Recognition

When the inputs take the form of images, the autoencoder demonstrates superior performance. Converting voice samples into images is achieved through the creation of spectrograms. As a result of this transformation, each voice sample is converted into a denoised and artifact-free spectrogram representation.[8]

This research experiment encompasses the utilization of various autoencoders. Training of these models has been conducted using speech samples from the aforementioned database. To facilitate training, testing, and validation, the database has been partitioned into distinct sections. Key metrics such as system accuracy, training loss, and validation loss have been computed for comprehensive evaluation of the system's performance.[9] The entire workflow, comprising these processes, is visually depicted in Figure 1.

Proposed Methodology

Database representation

All speech samples are transformed into spectrograms, as depicted in the block diagram, so that they can be used as input for training. The representation of the voice sample as a spectrogram is shown in Figure 2.

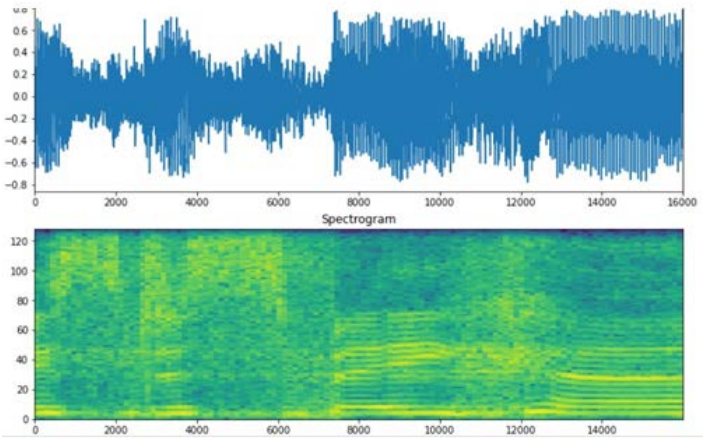
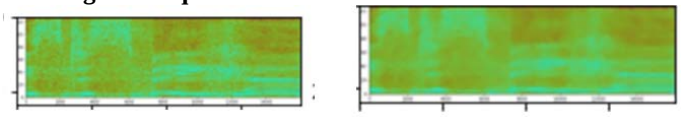


Figure 2. Voice sample representation as a Spectrogram

Activation function selection

The activation function is crucial to the efficiency of the system. Its purpose is to startle the cells into nonlinearity. The activation function will determine whether neuron cells will participate. In making judgments, it is therefore crucial. There are numerous activation mechanisms, such the sigmoid. However, we have used the Rectified Linear Unit (Relu) activation function in the case of our suggested model.[10]

Filtering Technique used



(a) (Before) (b)(after)

Figure 3. Spectrogram (after application of morphological filter)

The morphological filter's principle is the shrink and let grow method. The term "shrink" refers to the process of rounding off huge structures and removing small ones using a median filter before growing back the remaining structures by the same amount.

Instead of the coefficient matrix used in the linear filter, each matrix element in the morphological filter is referred to as a "structuring element." Only the values 0 and 1 are present in the structural elements. The dark shade component of the filter is also its hotspot.[11] The two-dimensional coordinate point sets that make up a binary image are described. Point Set is the term for this. The coordinate pair p = (u,v) of all foreground pixels makes up Q and

point set. Some point set operations are comparable to those in other images. Complement operation is used to invert binary images, and the union operator is used to combine two binary images. adding a vector to point p and shifting binary image I by some coordinate vector d. Or by multiplying -1 to point p, the binary image I can be reflected.[12] Figure 3. represents the spectrograms, after applying the morphological filters. (a) indicates spectrogram before filtering and (b) indicates spectrogram after application of a filter. Here, in this work we have considered input speech signals with background noises. Hence, if we apply morphological filters on noisy signal spectrograms, we can provide better input to the autoencoder.

Autoencoder

In this work we have used convolutional autoencoder, LSTM autoencoder and denoising autoencoder. Figure 4 shows the general model of Autoencoder system. It has two parts. Encoder section and decoder section. Latents features are the features which are present at bottleneck of autoencoder. In this case, the input to the autoencoder is the cleaned spectrograms by morphological filters of the speech signals.[13]

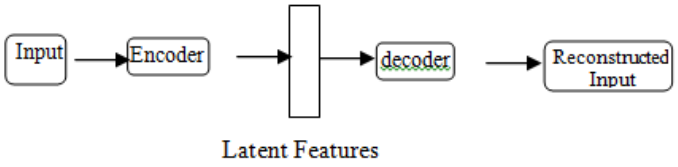


Figure 4. Structure of Autoencoder

Typically, the encoder can be expressed as a function g that depends on a number of variables.

$$h_i = g(x_i)$$

where, $h_i \in \mathbb{R}^q$, is output of the encoder section. when we test it against the input x_i, but, we have $g : \mathbb{R}^n \longrightarrow \mathbb{R}^q$

The Decoder section can be given as, $\sim x_i = t(h_i) - f''_3(x_i))$ (1)

Where, $\sim x_i \in \mathbb{R}^n$

Next step is training of autoencoder, means to satisfy g(.) and f(.) using following equation.

$$\arg\min (f,g) < [\ \Delta(\ x_i\ ,\ f(g(x_i))] > \qquad (2)$$

Where, Δ indicated the difference between input and output, which helps to decide the loss function. < . > indicates the average value of the observation.[1] It needs to find values of f and g such that, the autoencoder can reconstruct the successful output. Bottleneck is very important part of the autoencoder. It can be created using lower dimensions of features than the input features. So, basically input features plays very important role in the autoencoder functioning. Hence, in this proposed work, we have used morphological filter on the input data, so that we can provide clean features to the model, for testing and training purpose.

Experimental Analysis

Dataset used

This study employed a proprietary dataset consisting of a total of 50 voice samples. These samples were obtained from five distinct speakers, with a combination of utterances lasting either 3 seconds or 10 seconds. The dataset encompasses diverse texts used for both training and testing, and it also encompasses a range of languages to ensure its comprehensiveness.

The primary objective of our work is to develop a speaker recognition model. In pursuit of this goal, we extract and analyze the unique vocal attributes of each speaker. It's worth noting that the outcomes of analysis remain unaffected by the language used in the utterances.

Distinctive characteristics inherent to each individual speaker allow to perform accurate classification and identification tasks. Through this process, we can effectively differentiate and recognize speakers based on their specific vocal traits.[14]

Testing and training of the modeling

For testing and training processes, an augmented dataset is employed. We've employed augmentation techniques to effectively increase the size of the original database. Within this dataset, we've divided the samples into three key components: the training dataset, testing dataset, and validation dataset.

While the same dataset has been utilized for both testing and training purposes, our experimental design involves evaluating various scenarios. These scenarios encompass matching and mismatching conditions. A "matched" condition is defined when both the training and testing utterances possess identical durations. Conversely, a "mismatched" condition arises when there's a difference in the duration of utterances between training and testing phases.[15]

This experimentation incorporates three distinct types of autoencoders: convolutional autoencoder, LSTM autoencoder, and denoising autoencoder. Despite sharing a fundamental operational principle, these autoencoders exhibit unique characteristics. Figure 4 provides a visual representation of the Comparative Success Rate, showcasing the comparative outcomes obtained from these different autoencoder models.

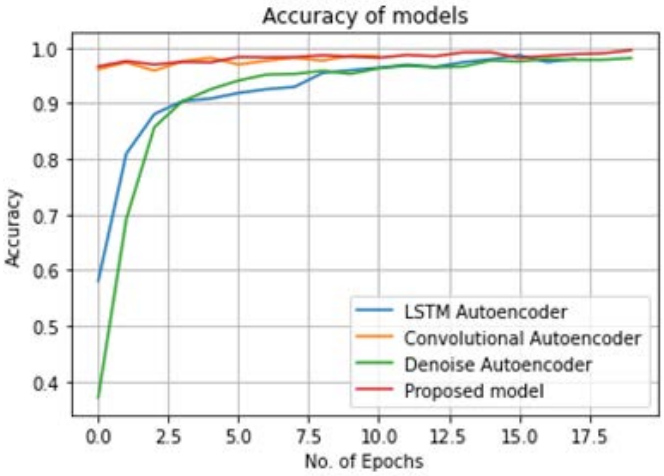


Figure 5. Comparative Success Rate of Autoencoders(in matched condition)

Whereas, figure 6 indicates the comparative loss of above mentioned autoencoders.

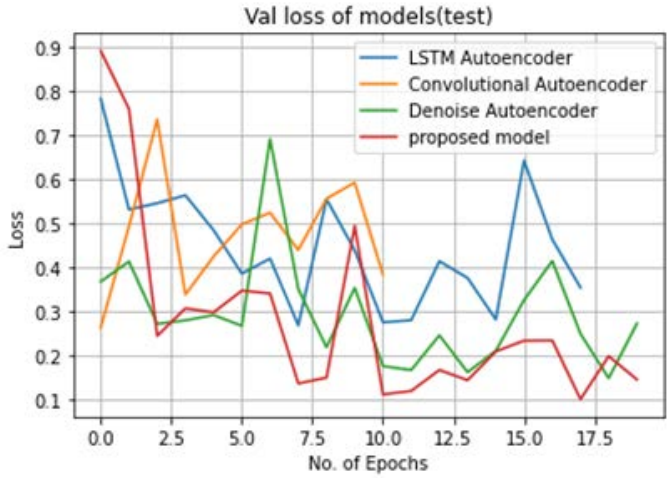


Figure 6. Comparative Loss Rate of Autoencoders (in matched condition)

Also, in order to validate our results, with the same dataset, we performed experiment with traditional SVM method and Random Forest method. The comparative analysis is given in Table 1. From the Table1. It shows that our proposed method is giving the better results in noisy environment. Our proposed method, not only gave result in clean speech environment, but also it has great results in noisy environment.[16,17]

Table 1. Comparison table of proposed system with existing system

Method	AUC	CA	F1	Precision
SVM	0.785	0.877	0.839	0.815
Random Forest	0.933	0.853	0.811	0.796
Our Model	**0.979**	**0.968**	**0.981**	**0.980**

Result Analysis

This study introduces an innovative speaker recognition system based on a convolutional autoencoder architecture. The system demonstrates commendable performance under matched utterance conditions, displaying a satisfactory success rate. However, its performance diminishes noticeably in cases where utterances are not temporally aligned. To address this limitation, we conducted experiments with various activation functions, observing significant improvements when employing the ReLU activation function.

Our approach involves utilizing raw voice samples without any preprocessing, which grants our system a degree of robustness against background noise. The experiments highlight that the primary challenge arises from the mismatched conditions between training and testing utterances. This discrepancy serves as a potential avenue for future research focus.

Furthermore, the inclusion of voice samples with background noise prompts the exploration of techniques to effectively eliminate these disturbances. This would enable the system to operate with a cleaner input database. The core of our proposed model leverages a convolutional autoencoder enhanced by a morphological filter, designed to mitigate background noise. It's important to note that no additional noises were deliberately introduced into the dataset.

The comparative assessment encompasses various autoencoder variants, including the denoise autoencoder, a convolutional autoencoder (excluding the morphological filter), and an LSTM autoencoder.[18] Additionally, we benchmarked our model against existing classification systems such as SVM and RF, further validating the effectiveness of our autoencoder-based approach. The results, as depicted in Figures 4 and 5, showcase a remarkable accuracy rate of approximately 98%.

To offer a comprehensive evaluation, we computed various performance parameters for our system, as summarized in Table 1. These results collectively underscore the robustness and efficacy of our proposed speaker recognition system

Conclusion and Future Scope

This endeavor focuses on enhancing the accuracy of a Speaker Recognition system. A pivotal element of our approach involves the utilization of a morphological filter to refine speech spectrograms, which are subsequently employed for training purposes. This filter proves effective in reducing noise within the image domain, contributing to enhanced data quality. Our process involves the conversion of speech signals into image representations through the creation of spectrograms. These spectrograms serve as visual depictions of the underlying speech signals.

In summary, our approach has yielded an impressive accuracy rate of approximately 98%. We also conducted comprehensive comparisons with various autoencoder models as well as traditional systems, providing a well-rounded assessment of our proposed system's performance. Future investigations could delve into addressing the challenge posed by mismatched conditions. In this study, we exclusively focused on matched conditions, where the duration of training and testing samples align. In contrast, mismatched conditions involve samples with differing durations. Additionally, researchers may explore alternative techniques for dataset cleansing, as dataset quality significantly impacts system accuracy. Enhancements to the dataset can consequently lead to increased system performance. In the broader context, the efficacy of our model underscores the critical role of the dataset in the experimental process. By directing attention towards optimizing the dataset, researchers have the potential to further elevate the system's accuracy.

Acknowledgments

Authors acknowledge the facilities provided by the university for carrying out this work.

Conflict of Interest Statement

Authors declare that there is no CoI for publication of this work.

References and Notes

1. P. Campbell. Speaker recognition: A tutorial. *Proc. IEEE* 85 (9), 1437–1462.

2. D. Gutman, Y. Bistritz. Speaker verification using phoneme-adapted Gaussian Mixture Models. *Eur. Signal Process. Conf.* **2002**, 2002-March (1), 19–41.
3. U. Michelucci. An Introduction to Autoencoders. **2022**.
4. L. Loina. Speaker Identification Using Small Artificial Neural Network on Small Dataset. In *Proceedings of International Conference on Smart Systems and Technologies, SST 2022*; Osijek, Croatia, **2022**; pp 141–145.
5. S. Hourri, N.S. Nikolov, J. Kharroubi. Convolutional neural network vectors for speaker recognition. *Int. J. Speech Technol.* **2021**, 24 (2), 389–400.
6. S. Kadyrov, C. Turan, A. Amirzhanov, C. Ozdemir. Speaker Recognition from Spectrogram Images. In *SIST 2021 - 2021 IEEE International Conference on Smart Information Systems and Technologies*; Kazakhstan, **2021**; pp 1–4.
7. K.C. Kishan, Z. Tan, L. Chen, et al. Openfeat: Improving Speaker Identification By Open-Set Few-Shot Embedding Adaptation With Transformer. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*; Singapore, Singapore, **2022**; Vol. 2022-May, pp 7062–7066.
8. E. Avots, T. Sapiński, M. Bachmann, D. Kamińska. Audiovisual emotion recognition in wild. *Mach. Vis. Appl.* **2019**, 30 (5), 975–985.
9. D. Nagajyothi, P. Siddaiah. Speech recognition using convolutional neural networks. *Int. J. Eng. Technol.* **2018**, 7 (4.6 Special Issue 6), 133–137.
10. S.S. Tirumala, S.R. Shahamiri, A.S. Garhwal, R. Wang. Speaker identification features extraction methods: A systematic review. In *Expert Systems with Applications*; **2017**; Vol. 90, pp 250–271.
11. Z. Liu, Z. Wu, T. Li, J. Li, C. Shen. GMM and CNN Hybrid Method for Short Utterance Speaker Recognition. *IEEE Trans. Ind. Informatics* **2018**, 14 (7), 3244–3252.
12. T. Kinnunen, H. Li. An overview of text-independent speaker recognition: From features to supervectors. *Speech Commun.* **2010**, 52 (1), 12–40.
13. J. Zhai, S. Zhang, J. Chen, Q. He. Autoencoder and Its Various Variants. In *Proceedings - 2018 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2018*; Miyazaki, Japan, **2019**; pp 415–419.
14. X. Feng, Y. Zhang, J. Glass. Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*; Florence, Italy, **2014**; pp 1759–1763.
15. R. Giri, M.L. Seltzer, J. Droppo, D. Yu. Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*; South Brisbane, QLD, Australia, **2015**; Vol. 2015-August, pp 5014–5018.
16. Y. Gao, J.X. Liu, L. Wang, J. Dang. Domain-adversarial autoencoder with attention based feature level fusion for speech emotion recognition. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*; Toronto, ON, Canada, **2021**; Vol. 2021-June, pp 6314–6318.
17. J. Rong, Y.P.P. Chen, M. Chowdhury, G. Li. Acoustic features extraction for emotion recognition. In *Proceedings - 6th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2007; 1st IEEE/ACIS International Workshop on e-Activity, IWEA 2007*; Melbourne, VIC, Australia, **2007**; pp 419–424.
18. S. Borde, V. Ratnaparkhe. Optimization in channel selection for EEG signal analysis of Sleep Disorder subjects. *J. Integr. Sci. Technol.* **2023**, 11 (3), 527.