Deep Learning based Voice assistance in hospitals using Face Recognition

G Jenifa¹, N Yuvaraj¹, B Karthikeyan² and K R Sri Preethaa¹

¹Department of Artificial Intelligence and Data Science, KPR Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India
²Department of Computer Science and Engineering, KPR Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India
jenifa.g@kpriet.ac.in

Abstract. After listening to a wake word or order, voice assistants come in quite compact packages and can accomplish a range of acts. They are able to switch on the lights, answering basic questions, playing music, place online orders, etc. As voice assistants become more robust, their usefulness can also be extended in both the personal and business areas. Various descriptive variables are fused in speech signals, leading to considerable difficulties interpreting any of the variables. There are various effective algorithms for speech and face recognition for the real-world applications. When a human try to communicate with a bot or vice-versa, there occurs so many difficulties in information sharing. This paper deals with one such application of bot assistance in hospitals, where a patient communicates with the bot assistant. Though the task seems to be quite easy, it involves a main threat: surrounding noise, word error rate, accuracy of the speech. An obvious impression is to factorize the frame of audio into various enlightening variables, but it turns out to be extremely difficult. The face of the patient who enters the healthcare premise is captured and analysed through the Deep Neural Network (DNN) based face recognition algorithm. Once identifying the user, the information regarding the patient is gathered from the database and given as a voice over output. Now, it is necessary to measure the accuracy of the word recognition and the word error rate. For this Cascade Deep Factorization (CDF) and Iterative Signal Enhancement (ISE) algorithm are used and discussed in this paper. For several speech processing functions like speaker recognition, this factorization and reconstruction method offers possible values.

1. Introduction

Speech Recognition is a mechanism by which an audio signal is translated into a text type, in which the voice recognition software works to identify the correct term that resembles each voice word. Speech is one of the most important resources for human and environmental communication, so the development of Automatic System Recognition is a dream for human all the time. The training process plays a very essential part in Speech Recognition. The design of a successful speech pattern training model increases the characteristics of overall success in the recognition of speech utterance. One of the most important biometric distinguishing strategies is facial recognition. It has a fair prospect of implementation in the areas of criminal recognition, device registration, management of household registrations, e-commerce, access control, etc.

Face recognition also has some problems in uncontrolled cases. Light, for example, has a major effect on the routine of face recognition systems, which degrades when the lighting is dim or dark [1].
There are many technologies has been revealed to address this problem and one among them is Multispectral imaging [2], and it is widely used in a variety of applications, like remote sensing applications [4], agricultural applications and food science related applications [3], medicine [5] and so on. The aim of multi-spectral images is to yield the advantage of the benefits of manifold spectral bands to show extra details than just one [6]. According to [7] and [8], the multispectral face identification system gives many advantages that includes: 1) It may show discriminating patterns in faces that colour trichromatic or monochromatic cameras are unable to capture [9]. 2) It permits for the intrinsic distinction of spectral disseminations between diverse face images, rendering them resistant to changes in illumination [10]. (3) The spectral factor, which extends the face space, as well as the spectral properties of different types of facial tissues, assist in improving individual discrimination [11].

Rich data, including linguistic content, speaker characteristic, mood, channel and background noise, etc., are included in speech signals. For many decades, researchers have worked to decipher this information, which lead to a multitude of tasks for processing speech information, that includes automatic speech recognition and speaker recognition [12]. Certain tasks have been solved very well after long-term studies, at least where a sufficient quantity of statistics is existing; while other methods are still challenging for example automatic emotion method [13].

2. Experimental study

This section defines the methodology proposed to develop the voice assistance system in hospitals using face recognition. Section 3.1.1 details about the proposed methodology for recognizing the face, retrieve the data of the identified person and communicate with the patient. Section 3.1.2 and 3.1.3 details the methods for speech recognition and identifying the word error rate.

2.1. Methodology

In the Big Data environment, local features extracted by the LBP algorithm have a stronger recognition effect. Local characteristics are simple to extract and use for identification, and computation time is cut in half. Integrating different local characteristics can also improve face recognition accuracy. LBP is a strong texture operator that marks image pixels by thresholding each pixel neighbourhood and treating the result as a binary number. Using histograms, LBP blends and denotes face images with the data vector. The LBP operator is then used to describe the relationship between each pixel in an image and its pixel value in the surrounding area.

2.1.1. Local feature extraction using LBP algorithm. In its most basic form, the LBP function vector is generated as follows:
1. Separate the window to be analysed into cells.
2. Each pixel in a cell is compared with each of its eight neighbours. This comparison takes place in a circular way, either clockwise or anti-clockwise direction.
3. When the centre pixel's value is higher than the neighbour’s value, write "0". Otherwise, form "1". This requires an 8-digit binary number.
4. Construct a histogram showing the frequency of "digit" in the cell. This histogram will appear as a 256-dimensional vector of features.
5. If needed, normalise the histogram.
6. Enter the histograms of all cells together (normalized). This produces a function vector for the entire window.

The patient's face is photographed before he or she arrives at the hospital. The face is recognised using the LBP with HOG algorithm. The backend extracts information such as the recognised face's name, age, last visit date, intention, and reason for visit. Using the extracted data, the user is given the patient's previous record as a text display on the computer or as a voice-over message. Along with the
chart, the patient is given instructions to follow when inside the hospital, as well as directions to see the doctor.

2.1.2. Algorithm for Iterative Signal Enhancement. Figure 1 shows the ISE procedure, which is still relatively new in the area of voice recognition, is based on a grouping of the input signal's SVD and an iterative aspect process that abstracts the signal's most active spectral properties each time. Frame by frame, the signal is interpreted, and an SVD of the i-th order is computed for each frame. The decomposition of the first-grade signal $x[n]$ through the distorted signal $z[n]$ is the first step in the algorithm. This scheming is achieved by trimming the Hankel matrix to grade one and calculating the average of anti-diagonals. The outstanding signal $o[n]$ is then determined by deducting the signal disintegration $x[n]$ through the raw noisy input $z[n]$. The method is iterated, using the outstanding signal as feedback for the upcoming iteration. When the residual signal's energy equals the noise's energy, the algorithm stops. The enriched signal $x[n]$ is attained by collecting the signal decomposition $x[n]$ over all repetitions.

![Figure 1. Iterative Signal Enhancement.](image)

2.1.3. Algorithm for Cascaded Deep Factorization. ASR research has shown that a DNN can infer linguistic content at the level of the frame independently [14], and the deep learning based utterer feature training method mentioned earlier in this work has shown that a very short segment can recognise speaker characteristics as well. Person deep factorization based on deep neural models denotes this single factor inference method (IDF). The IDF approach makes sense for two reasons: first, the end factor (language text or speaker characteristic) is satisfactorily relevant in speech signals; and second, there is a huge amount of training information existing. It's wide range supervised training that prefers the furthest task-oriented variables by holding the power of DNNs in attribute training [15]. IDF, on the other hand, isn't true for variables that aren't as important and/or don't have a lot of training data. Fortunately, good inference of language and speaker factors can greatly simplify the implication of the factors which are not happening in common in voice data. This inspired a cascaded deep factorization (CDF) method, which derives a unique IDF factor first, then make use of that factor as a conditional variable to derive the next factor, and so on. At the end, the voice signal will be factored with a variety of distinct variables, each of which corresponds to a specific task. WER = $s+d+(i/n)$ is a method for calculating the word error rate, where ‘$s$’ stands for the number of substitutions, ‘$d$’ stands for the number of deletions, ‘$i$’ equals the number of insertions and ‘$n$’ represents the number of words in a sentence.

3. Results and discussion

The communication between the bot and the patient involves much defined set of sentences like disease names, medication, patient names and other connective words only. Though the words involved are
definite, still there is possibility of error due to the factors such as surrounding noise, length of the word. The algorithms such as Iterative Signal Enhancement and Cascaded Deep Factorization are proved to be an effective algorithm in enhancing the speech quality. Hence, we used these two algorithms for our application and their respective word error rate (WER) with respect to the length of the word and the time length of the information shared by the bot with the patient. Table 1 and table 2 describes the WER percentage obtained for the prescribed data used in our application with respect to the length of the audio (in seconds) and the number of words in a sentence respectively using the ISE and CDF algorithms.

The data described in table 1 and table 2 is represented as a graph in figure 2 and figure 3 respectively. It shows that word error rate percentage is higher in ISE than that of the CDF algorithm irrespective of the number of words and the increase in the length of the audio.

**Table 1.** Word error rate (%) for the sentence with respect to the length of the audio.

| Time (in seconds) | ISE | CDF |
|------------------|-----|-----|
| 10               | 97  | 96  |
| 20               | 96  | 87  |
| 30               | 93  | 84  |
| 40               | 91  | 80  |
| 50               | 88  | 77  |
| 60               | 85  | 72  |
| 70               | 81  | 69  |

**Table 2.** Word error rate (%) for system voice with respect to the number of words in the audio.

| Length of sentence | ISE | CDF |
|--------------------|-----|-----|
| 23                 | 100 | 100 |
| 59                 | 97  | 92  |
| 85                 | 94  | 89  |
| 136                | 93  | 87  |
| 213                | 90  | 84  |
| 578                | 88  | 83  |
| 1026               | 87  | 79  |

**Figure 2.** Comparison of WER occurrence with respect to the length of the audio using ISE and CDF algorithm.
Various algorithms such as ISE and CDF have been applied and its performance has been monitored. Based on the performance evaluation, it has been observed that ISE provides the better efficiency.

4. Conclusion
In future voice assistance is going to be the main mode of communication between the machines and the humans. If texting is used for communication with the machine, it takes long time for conversation. Also, when native old people and unschooled people visit the hospital, it is difficult for them to text and they are unaware of the medical terms involved in hospitals. Thus, it is necessary to use an alternate and effective method such as voice can be used as input / output to avoid the complexities. These simple methods should also be an efficient one. So various algorithms such as ISE and CDF have been applied and its performance has been monitored for the application of healthcare assistance. The factors such as recognition accuracy and speed can also be measured using the same and different algorithms for this healthcare application as the words involved in the application is having the difficulty in pronunciation.

References
[1] Fei Wu, Xiao-Yuan Jing, Yujian Feng, Yi-mu Ji and Ruchuan Wang 2020 Spectrum-aware discriminative deep feature learning for multi-spectral face recognition, Pattern Recognition, Elsevier.
[2] Xiao Han and Qingdong D 2018 Research on Face Recognition Based on Deep Learning, IEEE.
[3] Yinhui Zhu and Yuzhen Jiang 2020 Optimization of face recognition algorithm based on deep learning multifeatured fusion driven by big data, Image and Vision Computing, Elsevier.
[4] Guodong Guo and Na Zhang 2019 A survey on deep learning based face recognition, Computer Vision and Image Understanding, Elsevier.
[5] Timo Ahonen, Abdenour Hadid and Matti Pietikäinen 2006 Face Description with Local Binary Patterns: Application to Face Recognition, IEEE transactions on pattern analysis and machine intelligence 28
[6] Nasir Saleem and Muhammad Irfan Khattak 2019 A review of supervised learning algorithms for single channel speech enhancement, International Journal of Speech Technology, Springer
[7] Youhao Yu 2012 Research on Speech Recognition Technology and Its Application, IEEE International Conference on Computer Science and Electronics Engineering.
[8] Haldorai, A. Ramu, and S. Murugan, Social Aware Cognitive Radio Networks, Social Network Analytics for Contemporary Business Organizations, pp. 188–202. doi:10.4018/978-1-5225-5097-6.ch010

[9] R. Arulmurugan and H. Anandakumar, Region-based seed point cell segmentation and detection for biomedical image analysis, International Journal of Biomedical Engineering and Technology, vol. 27, no. 4, p. 273, 2018.Zhao Xiu-Ying, Wang Hong-Yu, Tong Shou-yu, Fu De-you and Zhou Hai-shen 2012 Nonlinear Spectral Subtraction Method for elimination of Aircraft Engine’s Noise from degraded Speech Signals, Applied Mechanics and Materials.

[10] Ted H Applebaum and Brian A Hanson 1991 Regression Features for Recognition of Speech in Quiet and in Noise, IEEE.

[11] Afsaneh Asaei, Milos Cernak and Hervé Bourlard 2017 Perceptual Information Loss due to Impaired Speech Production, ACM Transactions on Audio, Speech, and Language Processing

[12] Theologos Athanaselis, Stelios Bakamidis, George Giannopoulos, Ioannis Dologlou and Evita Fotinea 2008 Robust Speech Recognition in the presence of noise using medical data, IEEE International Workshop on Imaging Systems and Techniques.

[13] Lantian Li, Dong Wang, Yixiang Chen, Ying Shi, Zhiyuan Tang and Thomas Fang Zheng 2018 Deep Factorization for Speech Signal, IEEE.

[14] Li Deng and Xiao Li 2013 Machine Learning Paradigms for Speech Recognition: An Overview, IEEE Transactions on Audio, Speech, and Language Processing