Retraction

Retraction: Voice Activated Face Recognition based Smart Support System (J. Phys.: Conf. Ser. 1916 012158)

Published 23 February 2022

This article (and all articles in the proceedings volume relating to the same conference) has been retracted by IOP Publishing following an extensive investigation in line with the COPE guidelines. This investigation has uncovered evidence of systematic manipulation of the publication process and considerable citation manipulation.

IOP Publishing respectfully requests that readers consider all work within this volume potentially unreliable, as the volume has not been through a credible peer review process.

IOP Publishing regrets that our usual quality checks did not identify these issues before publication, and have since put additional measures in place to try to prevent these issues from reoccurring. IOP Publishing wishes to credit anonymous whistleblowers and the Problematic Paper Screener [1] for bringing some of the above issues to our attention, prompting us to investigate further.

[1] Cabanac G, Labbé C and Magazinov A 2021 arXiv:2107.06751v1

Retraction published: 23 February 2022
Voice Activated Face Recognition based Smart Support System

N Yuvaraj¹, M Sanjeev², G Jenifa² and K R Sri Preethaa²
¹Student, Department of Computer Science and Engineering, KPR Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India
²Department of Artificial Intelligence and Data Science, KPR Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India
mysanjeev99@gmail.com

Abstract. Voice assistants usually respond to voice commands and provide the user with specific details about his question. Currently, voice assistants can process product orders, answer questions, perform tasks such as playing music, or initiate a quick phone call with a friend. For voice assistants, the long-term goal is to serve as a smart bridge between humans and the immense information and capabilities delivered by the Internet. Taking away the need to use some gadget or screen to communicate in various locations with the internet, technology or other humans. This paper deals with such a voice assistance for schedule maintenance for individuals which starts with the face recognition. Once a person enters the classroom / office, their face is captured and the face identification is done. If the identified person is the authorised one, then the system responds with a greeting message. When the person starts to ask about the schedule to the system, it responds with the schedule of the day of that identified person. In case of emergency schedules like meetings, the intimation is sent to the user as a message to alert him/her.

1. Introduction
A variety of approaches are used for auto-recognition speech systems (ASR), such as GMM-HMM and acoustic models based on the Deep Neural Network [1]. Recently there has been major breaches in the methods of end-to-end speech recognition [2] [3]. While several advancements were made with these ASR methods on clean voice signals, efficiency in the noisy and reverberation environment could be significantly degraded. In realistic settings, different background sounds and reverberations are often interfered with captured speech signals. It is therefore very important to improve the sturdiness of auto-recognition of the speech systems. The goal of this paper is to compare the peer-to-peer speech recognition's noise robustness.

To improve the strength of Automatic Speech Recognition, there are three mainstream methods, including way of enhancing the automatic speech recognition by considering the surrounding noise as well. The first mainstream solution is to introduce the voice improvement portion at ASR's front end.

Methods of speech improvement include spectral subtraction [4], Wiener filtering [5] and speech enhancement [6] [7] based on the DNN. However, improving speech optimizes the models in order to evaluate the end voice that differs from the voice identification portion. The methods used to enhance the speech therefore struggle to maximize the overall target, leading to a sub-optimal solution [8].

Furthermore, the improved speech produced by these methods of speech enhancement produces extra polished speech, which is the explanation for speech misrepresentation after enhancing the speech. The
misrepresentation of speech will reduce the result of ASR [9]. The efficiency of this method is therefore extremely reliant on the routine of the front-end enhancement [10]. In order to recover the noise sturdiness of ASR, the second conventional approach uses multicondition training. To train the models of speech recognition, MCT utilizes various types of statistics (clean and loud speech). The difficulty and computational outlays of MCT have, however, enlarged. In addition, on the unmatched conditions [11], it provides unimpressive performance and the speech distortion [12] also affects the performance.

To discourse the voice alteration issues, the enrichment of front-end improves initially both the training and sample sets and ASR voice model is then trained with the enriched voice data. It can boost the efficiency of the ASR to certain extent, but still, it relies heavily on the routine of the front-end enhancement. The SpecAugment [13] explicitly applies the increase in data to the input features of neural networks, unlike the MCT process. The preparation involves three spectrogram distortions such as time warping, time and frequency masking. Although the SpecAugment can enhance the peer-to-peer ASR performance, the noisy situation needs to be improved.

Joint training methods [14], [15] are the third form of mainstream. Such approaches apply this system to simultaneously optimize the improvement and understanding of voice. The explanation is that speech improvement and recognition of speech are not two different activities and they can benefit obviously from each other. A collaborative adversarial enhancement training method was employed in order to improve the noise sturdiness of peer-to-peer ASR [16].

2. Proposed methodology

Depending on the types of utterances to be understood, various types of speech recognition systems are available. The following are the classifications for these different types of utterances:

**Isolated Words**

Isolated word recognizers usually achieve silence on both sides of the sample window for each utterance. There are typically two states of these schemes. When the speaker must pause between two utterances, he or she is in a Listen/Not-Listen condition. Speech signals are processed during the pauses between utterances.

**Words That Are Connected**

Related words are identical to isolated words, with the exception of a brief pause.

**Continuous Terms**

Continuous speech comprehension necessitates a nearly natural speaking style. Continuous speech recognizers are difficult to design because they require special methods to establish utterance boundaries.

**Words That Come Out of the Blue**

Spontaneous speech conceals mispronounced, unrehearsed non-words with difficult-to-read false claims [17]. This type of ASR framework is designed to manage a variety of features, such as terms that are run together, such as "ums" and "ahs" [18].

Speech recognition methodologies are generally categorised into three categories: acoustic–phonetic approach, pattern recognition approach, and artificial intelligence approach.

2.1. Using Artificial Intelligence

It's a hybrid approach that incorporates the principles of acoustic phonetics and pattern recognition. Dynamic time warping (DTW) and HMM are broadly used in pattern matching [19].

Speech recognition in DTW is focused on grades. One or more models may be used to represent each class [20]. The machine modelling improves as the number of templates rises. Hidden Markov Model (HMM) is favoured over DTW in advanced systems due to better performance and lower memory requirements. This method is used for complex tasks, but it is inefficient when working with large data sets. Artificial neural networks' basic approach is phoneme recognition. This is achieved using an intelligence strategy that includes interpreting and visualising the input expression. A large number of neurons make up the network. Each neuron calculates the nonlinear weight of inputs and sends the result
to the incoming units. This method of acquiring training sets assists in assigning values to input and output neurons.

![Image](image.png)

**Figure 1.** Speech Recognition.

### 2.2. Algorithm for Mel-frequency cepstral co-efficient

The frequency axis of the MFCC is primarily distorted to the mel scale value, that is approximately less than 3 kHz and logarithmic greater than this point. On the twisted continuum, a triangular filter with similar spacing in the mel-scale is applied. To attain MFCC feature vector for voice data that is the spoken words, the product of the filters is compressed using Log function and cepstral coefficient is calculated by using the DCT. MFCC is a signal acuity model that more closely approaches the human hearing system's reply than the usual spectrum's linearly-spaced frequency bands. The following is the MFCC algorithm's frame:

**Phase 1:** The Discrete Fourier Transform must be used to convert the input signal $y(k)$, that undergoes a difficult sequence of conversions in the beginning stages of auditive processing.

**Phase 2:** Using triangular overlapping windows, map the powers of the spectrum that goes greater than the mel-scale.

**Phase 3:** Measure the logs of powers for each mel-frequency.

**Phase 4:** Take the list of mel log powers and add the discrete cosine transform (DCT) to it.

**Phase 5:** The MFCCs are the resulting spectrum's amplitudes.

Figure 1 shows the steps involved in the speech recognition system with the feature extraction and recognizer phases.

### 3. Results and discussion

The outputs produced by the machine and the humans are always in different pitch because of the change in the pronunciation. The methods such as Mask-Based Enhancement Network (MBE) and Gated Recurrent Fusion Network (GRF) are used for identifying the variations in the voice messages. Figure 2.a, 2.b, 2.c shows the waveform of the voice produced by human to communicate with the machines and the length of the voice is 10, 20 and 30 seconds respectively and contains individual, connected and spontaneous words. Figure 3.a, 3.b, 3.c shows the waveform of the voice produced by machine to communicate with the humans while trying to answer their question and the length of the voice is 10, 20 and 30 seconds respectively with individual, connected and spontaneous words.
Figure 2. Waveform of Human Voice.

a. Waveform of human voice of length 10 seconds
b. Waveform of human voice of length 20 seconds
c. Waveform of human voice of length 30 seconds.

Figure 3. Waveform of system generated voice.

a. Waveform of system generated voice of length 10 seconds
b. Waveform of system generated voice of length 20 seconds
c. Waveform of system generated voice of length 30 seconds.

Figure 4.a, 4.b, and 4.c represents the spectrogram of the voice produced by human to communicate with the machines and the length of the voice is 10, 20 and 30 seconds respectively. Figure 5.a, 5.b, and 5.c represents the spectrogram of the voice produced by system to respond to the humans and the length of the voice is 10, 20 and 30 seconds respectively.
Figure 4. Spectrogram of human voice.

a. Spectrogram of human voice of length 10 seconds
b. Spectrogram of human voice of length 10 seconds voice of length 20 seconds
c. Spectrogram of human voice of length 10 seconds voice of length 30 seconds.

Figure 5. Spectrogram of system generated voice.

a. Spectrogram of system generated voice of length 10 seconds
b. Spectrogram of system generated voice of length 10 seconds voice of length 20 seconds
c. Spectrogram of system generated voice of length 10 seconds voice of length 30 seconds.

Figure 6.a, 6.b, and 6.c represents the log transformation of the voice produced by human to communicate with the machines and the length of the voice is 10, 20 and 30 seconds respectively. Figure 7.a, 7.b, and 7.c represents the log transformation of the voice produced by system to respond to the humans and the length of the voice is 10, 20 and 30 seconds respectively. The blue colour of the log transformation represents the high pitch of the voice and the orange colour represents the normal voice communication without the sudden modification in the voice. From these various representations of the voice, it can be seen that the change in the voice modulation of the human is higher than the machine
voice. The output voice by machine also gets affected by the external noise which complicates the understandability. Hence while developing applications which involves the voice mode of communication instead of tradition texting, it is necessary to concentrate on the voice modulations in machine voice so that the understanding by the human is good.

Figure 6. Log Transformation of Human Voice.

a. Log transformation of human voice of length 10 seconds. b. Log transformation of human voice of length 10 seconds voice of length 20 seconds. c. Log transformation of human voice of length 10 seconds voice of length 30 seconds.

Figure 7. Log Transformation of System Generated Voice.

a. Log transformation of system generated voice of length 10 seconds. b. Log transformation of system generated voice of length 10 seconds voice of length 20 seconds. c. Log transformation of system generated voice of length 10 seconds voice of length 30 seconds.

Various algorithms such as ISE and NSS has been applied and its performance has been monitored. Based on the performance evaluation, it has been observed that ISE provides the better efficiency.
Table 1. Percentage (%) of word / sentence recognition accuracy

| Type of Word | Recognition |
|--------------|-------------|
| Isolated     | 100         |
| Connected    | 98          |
| Continuous   | 96          |
| Spontaneous  | 93          |

Table 1 shows the percentage of accuracy with respect to the different kinds of words. The sentences included the length from 10 seconds to 30 seconds. It includes collection of isolated, connected, continuous and spontaneous words.

4. Conclusion
The voice communication is an effective type of knowledge transfer between the machines and humans. There is always difference between the human and the machine voice. It is a major concern when the transfer of data between the user and bot happens only through voice. So, this paper discuss about the various methods for speech recognition and the complexities arises when the machine talks. This paper deals with the situation when an application depends only with the voice mode of communication, the rate of understandability of the user. The application must deal the factors like external noise, recognition accuracy, speed of recognition which are the major source of misunderstanding. These factors differ from word to word depends on its length, type of pronunciation.

References
[1] Ramya Devi M and Dr Ravichandran T 2013 A Novel Approach for Speech Feature Extraction by Cubic-Log Compression in MFCC, proceedings of the International Conference on Pattern Recognition, Informatics and Mobile Engineering.
[2] Carlos Molina, Nestor Becerra Yoma, Fernando Huenupán, Claudio Garretón and Jorge Wuth 2010 Maximum Entropy-Based Reinforcement Learning Using a Confidence Measure in Speech Recognition for Telephone Speech, IEEE Transactions on Audio, Speech, And Language Processing 18.
[3] Radmio Peic 2013 A Speech Recognition Algorithm Based on The Features of Croatian Language, EURASIP Conference focused on Video Image Processing and Multimedia Communications.
[4] Cunhang Fan, Jianshan Yi, Jianhua Tao, Zhengkun Tian, Bin Liu and Zhengqi Wen 2021 Gated Recurrent Fusion with Joint Training Framework for Robust End-to-End Speech Recognition, IEEE/ACM Transactions on Audio, Speech, And Language Processing 29.
[5] Aswin Shanmugam Subramaniam, Xiaofei Wang, Murali Karthick Baskar, Shinji Watanabe, Toru Taniguchi, Dung Tran and Yuji Fujita 2019 Speech Enhancement Using End-To-End Speech Recognition Objectives, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.
[6] Trishna Barman and Nabamita Deb 2017 State of the Art Review of Speech Recognition using Genetic Algorithm, IEEE International Conference on Power, Control, Signals and Instrumentation Engineering.
[7] Orken Mamyrbayev, Nurbapa Mekebayev, Mussa Turdalyuly, Nurzhamal Oshanova, Tolga Ihsan Medeni and Algerim Yessentay 2019 Voice Identification Using Classification Algorithms, Intelligent System and Computing.
[8] H. Anandakumar and K. Umamaheswari, A bio-inspired swarm intelligence technique for social aware cognitive radio handovers, Computers & Electrical Engineering, vol. 71, pp. 925–937, Oct. 2018. doi:10.1016/j.compeleceng.2017.09.016
[9] R. Arulmurugan and H. Anandakumar, Early Detection of Lung Cancer Using Wavelet Feature Descriptor and Feed Forward Back Propagation Neural Networks Classifier, Lecture Notes in Computational Vision and Biomechanics, pp. 103–110, 2018. doi:10.1007/978-3-319-71767-8_9
[10] Xiao Han and Qingdong Du 2018 Research on Face Recognition Based on Deep Learning, IEEE.
[11] Yinghui 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.

[12] Guodong Guo and Na Zhang 2019 A survey on deep learning based face recognition, Computer Vision and Image Understanding, Elsevier.

[13] 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.

[14] Nasir Saleem and Muhammad Irfan Khattak 2019 A review of supervised learning algorithms for single channel speech enhancement, International Journal of Speech Technology, Springer.

[15] Youhao Yu 2012 Research on Speech Recognition Technology and Its Application, IEEE International Conference on Computer Science and Electronics Engineering.

[16] Ali Bou Nassif, Ismail Shahin, Imtinan Attili, Mohammad Azzeh and Khaled Shaalan 2019 Speech Recognition Using Deep Neural Networks: A Systematic Review, IEEE Translations and content mining.

[17] Vlado Delić, Zoran Perić, Milan Sećujski, Niksa Jakovljević, Jelena Nikolić, Dragiša Mišković, Nikola Simić, Siniša Suzić and Tijana Delić 2019 Speech Technology Progress Based on New Machine Learning Paradigm, Computational Intelligence and Neuroscience.

[18] 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.

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

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