FOREX Publication
Open Access | Rapid and quality publishing

International Journal of
Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 4 | Pages 862-867 | e-ISSN: 2347-470X

Speaker Recognition Assessment in a Continuous System for Speaker Identification

Mahesh K. Singh[1*], P. Mohana Satya[2], Vella Satyanarayana[3] and Sridevi Gamini[4]
[1,2,3,4]Department of ECE, Aditya Engineering College, Surampalem, India, [1]mahesh.singh@accendere.co.in, [2]perurimohanasatya999@gmail.com, [3]vasece_vella@aec.edu.in, [4]sridevi_gamini@yahoo.com

***Correspondence:** Mahesh K. Singh; mahesh.singh@accendere.co.in

ABSTRACT- This research article presented and focused on recognizing speakers through multi-speaker speeches. The participation of several speakers includes every conference, talk or discussion. This type of talk has different problems as well as stages of processing. Challenges include the unique impurity of the surroundings, the involvement of speakers, speaker distance, microphone equipment etc. In addition to addressing these hurdles in real time, there are also problems in the treatment of the multi-speaker speech. Identifying speech segments, separating the speaking segments, constructing clusters of similar segments and finally recognizing the speaker using these segments are the common sequential operations in the context of multi-speaker speech recognition. All linked phases of speech recognition processes are discussed with relevant methodologies in this article. This entire article will examine the common metrics, methods and conduct. This paper examined the algorithm of speech recognition system at different stages. The voice recognition systems are built through many phases such as voice filter, speaker segmentation, speaker idolization and the recognition of the speaker by 20 speakers.

Keywords: Speaker recognition, DSL, SVM, BPNN, Speaker identification.

ARTICLE INFORMATION
Author(s): Mahesh K. Singh, P. Mohana Satya, Vella Satyanarayana and Sridevi Gamini;
Received:28/06/2022; **Accepted**:01/10/2022; **Published**: 18/10/2022;
e-ISSN: 2347-470X;
Paper Id: IJEER-RDEC5;
Citation: 10.37391/IJEER.100418
Webpage-link:
https://ijeer.forexjournal.co.in/archive/volume-10/ijeer-100418.html

Publisher's Note: FOREX Publication stays neutral with regard to Jurisdictional claims in Published maps and institutional affiliations.

1. INTRODUCTION

expressive biometric functionality of different automated uses [1]. Speaking mechanism is the general phenomenon associated with identity detection, emotional recognition, attention recognitions, etc. Many technologies for globalized communication independently of language restrictions are used for speech processing. Speak processing is used for verifying the identity of the user as an integrated feature in different online and offline systems [2]. The automated answering machines often provide instant feedback using speech recognition techniques to communicate with clients and users. Both these implementations can be summed up as speech acknowledgement or speaker acknowledgement, or language transfer. Each of these systems is collectively recognized [4] as a pattern where speech signal is processed to respond to similar questions in different aspects. In any method of speech processing [1] both conventional and scientific approach must acquire and process the speech characteristics. The conventional phenomenon indicates that speech should be treated as dictations or word processing. The word isolation, word delay, pitch of spoken words are the common features used to take adaptive decisions by most speech processing systems. In the speaker systems, the precision of the language of the speaker, the speech disturbance and the ecosystem constraints are affected. The language-influenced acoustic methods [3] are often used to predict words and speech. There are also 2 technological issues and considerations in order to correctly and properly recognize the voice [5]. This article discusses different kinds of speech processing systems, different problems and common methods for generating features. This caption identifies and discusses each contributing factor which may affect the speech system.

1.1 Factors of Speech Variability

In pre-processing and feature generation phases, multi-layer speech processing is needed to address variability [2] in acquired voice signals. The structure and quality of the speaker signals can be affected by different technological, environmental and user-specific factors. These variability factors may recognize the individual, sex, conduct, custom or some other information hidden by users. These variability features must be suppressed or examined to make the right decisions in accordance with the application and specifications. Different factors are described and addressed in this section which can lead to variations in the speech signal at semitone level, language level and structure level [6].

1.2 Speaker Accent

Two speakers can vary due to their native and non-native speakers' physiological properties. Likewise, regardless of the sex of people, the pitch and rhythm of speech can be different [7]. A speaker's geographical variation can be substantially associated with the speaker's accent. Accent is determined by the mother tongue, skill level and acoustic switching. The native speaker-trained speech system could not produce good results for non-native speakers. The students also regard the accent classification as a separate field of study. The focus also has a

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 4 | Pages 862-867 | e-ISSN: 2347-470X

larger reach for global communication or inter-language communication [8].

1.3 Physiology Representation

The vocal aperture not only influences the geographic roots, but also the voice pitch, the flow and structural characteristics [9]. The speaker recognition systems are also highly used to track a speaker's presence. The error rates for a device based on a speaker are much higher than for individual speaker systems. The hole between the lips and glottis has an impact on the variations in the signal. It can affect the speaking time, the loudness of two words. Every speaker's vocal organ form is unique. In addition, frequencies and structural features of speech are recognized as glottal pulse and its effect on voice quality [10]. Researchers applied the frequency level assessment or the band pass assessment compensates certain differences. The variability between speakers can be reduced and/or increased, as per the application requirements, to improve the robustness of any speech processing device [11]. A different measure can be taken to the corresponding feature-assisted selective field. 27 can be merged into a single and combined form, so that the language or featured area can be effectively extracted. *Figure 1* shows the external view of the voice segmentation frame [12].

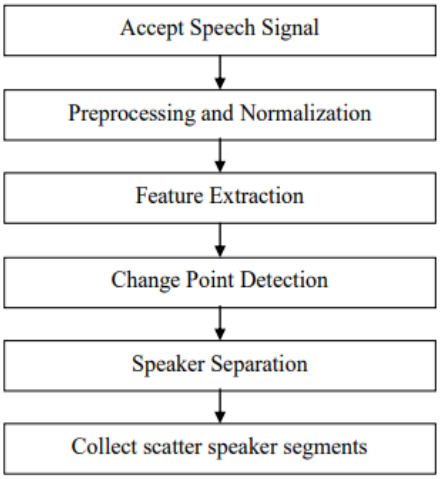

Figure 1: Speech Segmentation Process

The collective aspects of speech with associated segments are shown in *figure 1*. The segmentation is made here for extraction of pitch and the function area for automatic speech recognition (ASR) [13]. In speaker recognition systems, the pitch features are of better importance and reach. For speaker recognition the speed, phonetic rate, vowel time, pause time and other temporal adjustment features can be extracted [14].

2. SYSTEM MODEL

One of the most expressive elements in the relationship between human and machine is to understand or identify users' emotions. The emotions affect a person's face and voice. The voice can be thick, light or emotionally shaken. By conducting speech analysis these emotions can be identified automated. For an efficient recognition of language and isolation of related feelings, various kinds of speech features can be extracted. To identify these emotions from expression, researchers have identified diverse feature processing & classification methods. Emotions are common: colure, sorrow, terror, astonishment, happiness 42 and disgust. Research was carried out using important aspects and function observation of various emotional classification methods. [1, 15]. The author addressed the availability of speech emotion databases, methods of feature generation, and their categorization and extent in relation to emotion. Chen et al. had proposed the emotion recognition method at three levels and generated the characteristics at each level by using fishing rate [2].

The author under the SVM (Support Vector Machine) classifier processed these generated features for speech emotion recognition. The typical energy and frequency traits, the basic characteristics of the continuum and the specific characteristics of fishermen have been processed. The author suggested three-level architecture. The emotion set at every level isolated the same emotional class precisely in this architecture. In order to recognize the emotion from spontaneous expression, an HMM (Hidden Markov Model) probabilistic classifier has been defined [3, 16]. In the Frequency Vector, wavelets, energies and entropy were processed to classify speech emotion effectively. To predict the emotional class, a ranking and speech responsive method [4, 17] has been implemented. In order to understand the certain emotion, the basic intuitions were added. The neutral pronouncement is contrasted relatively with the classification characteristics and the emotion of speech is defined on the basis. To improve the exactness of the speech emotion prediction a Fourier based perceptual consent mapped was created [5, 18, 19].

SVM Classifier for speaker independent emotion recognition was used to process the best dynamic and continuous features. For the generation of structural features used double sparse learning (DSL) in order to achieve a promising recognition of speech emotion [6]. The author split the speech into smaller segments and created special characteristics for the super vector. The DSL method is used to produce weight of the function and to set weights on signal 43 segments in order to accurately classify the emotion. As subsequent vectors of features, low-level descriptors [7] are integrated. These features have been combined with short-term statistics, autoregressive models and spectral moments. These parameters were grouped using a wavelet to classify speech emotions efficiently. The method based on log-like features was used to identify emotion by using a hidden Markova model. For every speech frame, Yoon et al. created the basic pitch, energy and MFCC features [8]. Amplitude and magnitude analysis were carried out by the author with the noisy speech signal. These short-term sensory functions were formulated for the identification of the signal community and the emotion of certain speaker functions. Work was done with entropy and probabilistic methods for recognizing wrath [9]. In order to produce function score and calculate relative weights the author has processed the acoustic and linguistic features. For successful identification of the rage from the speech signal, different types of frequency features have been processed [19, 21, 22].

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 4 | Pages 862-867 | e-ISSN: 2347-470X

Stress is such a human function or action that influences the way speech is spoken. The processing of speech disturbs the tension of the natural speech. The voice can be louder, of higher pitch, more frequent, and quicker in the event of stress. In order to learn speech using data treatment and optimize speech via filter bank, it had proposed a method [13]. The population-based meta-heuristic approach was used to classify stress speech as an evolutionary algorithm. There was a hamming window for the acquisition of the frame features and for the separation of speech utterance for identification of tension. A Fourier spectrum for the 44 classifications of stressed speech [20, 23, 24] was used for a harmonic peak to energy ratio measurement. The characteristics are dependent on the amplitude assessment at breathability.

In conjunction with harmonic peaks, the statistical test characterizes various stress speech groups. To improve the classification results, the created features have been managed with binary cascaded SVM. The study of different methods of characteristic generations for improved recognition of stressed speech was conducted [18]. The study aimed at identifying the condition-dependent functionality with further compensation and adjustment. The speech style is defined such as rage, loudness and the consequences on speech. The author has examined the power spectrum and Fourier characteristics with ambient noise. The author also observed different parameters and their effect. The condition of trust is also under stress assessment [14]. The measure of trust can be predicted by means of speech analysis, speech delivery speed etc. The author described the speaker adapted model to analyze the lectures.

3. SPEECH RECOGNITION METHODOLOGY

The research work discussed focuses on the identification of speakers by multi-speakers. The presence of multiple speakers involves every meeting, interaction or debate. This type of speech presents different difficulties and stages of processing. The challenges include environmental impurities, speaker participation, speaker distance, microphone equipment, etc. In addition to handling these real time difficulties, there are certain process-level problems with multi-speaker speech processing. Identifying the speech segments, separating the speaking segments, creating the clusters of similar segments and then recognizing the speaker by observing them are the typical sequential processes in recognition by multi-speaker language. Both related phases of speaker recognition processes are examined with relative methods in this study. This entire article examines the typical actions, technique and attitudes of these work phases. This method of improving speech is characterized by the analytical action of the voice signal. Speech signals of this type need the encoding of speech. Figure 2 shows the subtractive algorithmic adaptive process. In the context of Fourier transformation, the communication models are described STFT. This type of speech signal processing is often performed with an amplitude analysis, so that the application accepts the noise inclusive signal. The estimate of this signal is done for an adaptive study of the frame.

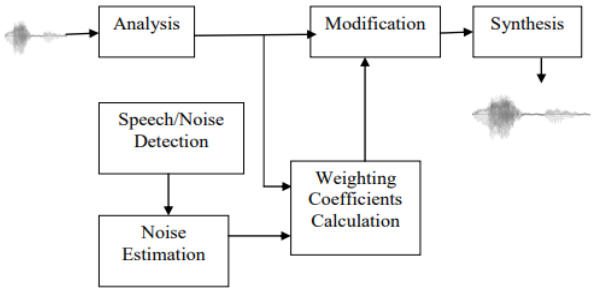

Figure 2: Speaker Recognition Algorithm

3.1 Speaker Recognition Methods

In biometric authentication systems, emotional systems and commitment assessment systems, the speaker recognition systems are adopted. The identification of the speaker can be implemented with or without mapping of the wording information. Two interconnected phenomena [19] classify speech recognition systems. These phenomena are stochastic approaches and models.

Stochastic Approach: This approach is aligned with the dictionary and the language specification. The study of the sentencing standard takes place by earlier setting of the parameters. The sentence-based comment is applied to map the content of the dictionary with voice elements. The procedure for identifying the voice elements or contents is used for maximum probability analysis. The likelihood assessment is Analysis Detection of talk/noise Estimation of noise Calculation of weighting factors Change Synthesis39 of voice components with quantity measurement calculation. The estimate by probability is used to map the spoken pronunciation in the sentence. This approach is also intended for the sequence observation. The standardized distribution of probabilities is calculated to estimate matching costs. For speech recognition, the complexity of alternative measures to probability measures is also applied. This approach is used to increase the matching rates in language and acoustic modeling.

Template Based Approach: This approach defines the smaller dictionaries and templates to store different language-based aspects such as contents, language-specific templates and quiet templates. For identification of speakers by prototype mapping, the maximum probability criterion is applied. The consistency of the corresponding template is determined by the size and number of templates. The special function, utterance and speaker specific models can also be used to increase the efficiency of a speaker recognition system. Various distance measurements are available to fit input speech to the reference modules, such as log defined distance, cepstral distance, probability distortion and weighted cepstral distance methods. In order to avoid any potential mistakes, the dictionary applications are described with small and large expressions. The Acoustic Dictionaries sensitive to the emotion or the speaking group are also used.

3.2 Data Set Description

Validated the proposed multi-speaker recognition model for Center for Speech Technology Research – Voice Cloning

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 4 | Pages 862-867 | e-ISSN: 2347-470X

Toolkit (CSTR-VCT). The full collection of data contains 200 native speakers recorded in English. The emphasis of these speakers can vary. The talk is recorded at a frequency of 80 KHz. Each speaker reads the recordings of the paper, the rainbow message and the phrase of an excitation for approximately 200 phrases. Table 1 describes the data set.

Table 1: Speaker's data set description for experiment

Important parameter	Respective values
Recording Device	A head-mounted microphone
Number of individual speakers sentences	200
Spoken language	English with Accent variation
Speech Recording Sampling Frequency	80
No. of speakers	200
Dataset Name	CSTR VCTK Corpus

Database characteristics produced by robustness consideration shown in *table 1*. Multiple speakers with different sentence segments contribute to a speech signal. There is a difference in the dialect and the content of speech. The segments of the speech are from various sources obtained. The results of the speech processing on speech instances from this data set are presented in this article. This article presents the analysis comparative findings for different data sets.

4. RESULTS AND DISCUSSION

The acquired real-time speech can be infected with content or the silence area in real-time. Probabilistic measures are implemented in order to turn speech into a standardized type and to increase the rate of recognition. In this portion, the results are given on the specimen speaker segment taking into account the segment size of 64ms. In the case of character speech processing, the comparatively better results for smaller segments are verified. This sub-section displays the input speech, the process result speech, and the standard function processing sentence speech shown in *figure 3*.

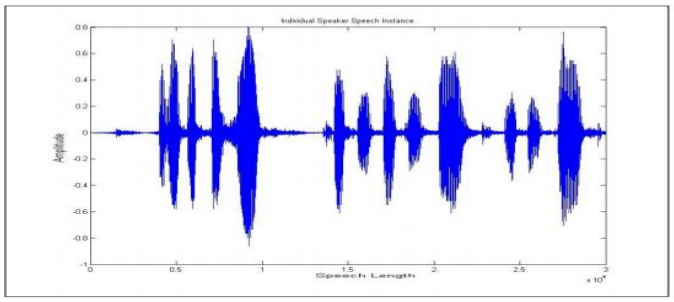

Figure 3: Input Speech Sample

The speech can be influenced by different noisy vectors of the atmosphere and system as it is gathered in the real environment. The phrase is more important towards the vector of noise. The input raw talk from the atmosphere is shown in this figure. As there can be many words in the voice, the words can be paused. The area of the content shows the actual data aspect to recognise the language. The sample talk is randomly picked from test speech instances. The correction can be used to filter the speech and increase the speech strength. In order to minimize noise and increase signal power, high and low-level filters are used sequentially. This manuscript provides speaker recognition as a sub-model for the individual speaker data set. The character and sentence speech data sets are recognized by the speaker. The character speech data set is manually collected using various microphone devices. The alphabets in English are recorded at the frequency 48 KHz. To create the data set, 10 male and 10 female speakers record 10 instances. This basic set of data is described in *table 2*.

Table 2: Speaker Recognition characteristics data set description

Parameter Setting	Values
Speech language	English
Type of dataset	Primary
Speakers	10 Female, 10 Male
Instance per speakers	20
Environment of speech capture	Office
Recording device	Microphone

In the case of continuous expression, which is supported by many speakers, the multidisciplinary recognition of speakers is achieved. The parts of the speech are generated by separating the speakers. The speaker count is marked by each instance of the speaker. Under the score adaptive deep neural network, the speaker featured data set is processed to perform the classification. The raw input speech is processed using different steps and algorithmic methods to process and recognize speakers. A specimen for intermediate work stage results is processed here. One speech sample. The speech is made by many speakers with various parts of the speech and different spaces. This segment provides the graphic presentation of these characteristics and speech characterization. The sample talk in *figure 4* has the least noise effect. *Figure 4* shows the contribution of several speakers and the relative frequency and pitch observation.

For specified characteristics, the suggested probabilistically weighted SVM model is used to recognize the speaker. The robust models are defined so that noise, regardless of the sex criteria, adversely affects language. Each test set has 60 and 100 occurrences of speech with noise variation and gender-specific speakers. *Table 2* provides a comparative assessment with regard to the mapped number of cases. PCA and BPNN classifiers are presented with analytical observations.

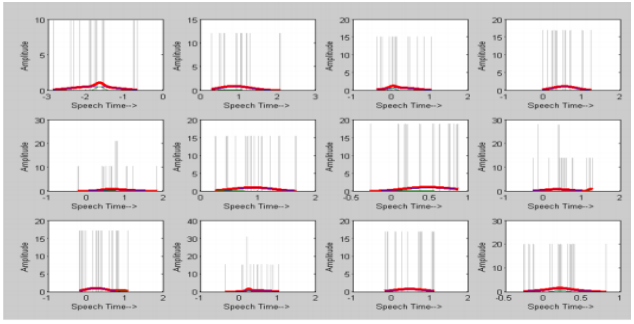

Figure 4: Estimation of Multi Speaker Contribution

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 4 | Pages 862-867 | e-ISSN: 2347-470X

Table 3: Speaker Recognition Rate Analysis

Speakers	Text Set Size	PCA	BPNN	SVM (Proposed)
S1	60	67%	63%	80%
S2	100	64%	70%	84%
S3	100	60%	56%	78%
S4	100	62%	67%	86%
S5	100	66%	59%	79%

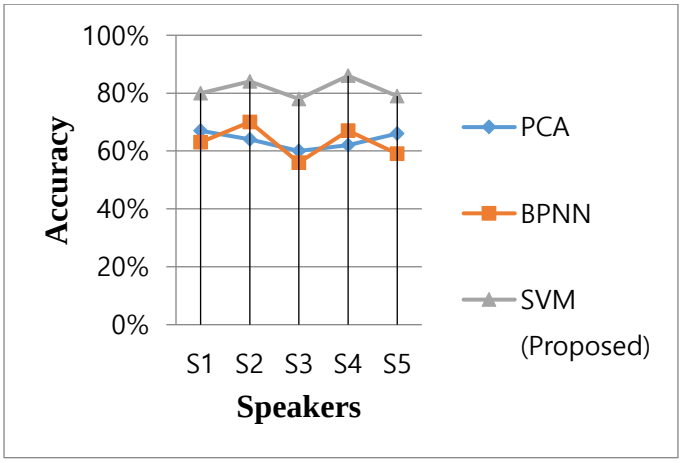

Figure 5: Characteristics of Specific Speaker Recognition

The character-specific speaker recognition is illustrated graphically in *figure 5*. The x-axis indicates the classificatory used to recognize the speaker independently by the speaker. The diagram shows all three samples of comparison results. For the precision ratio evaluation, the analytical findings are obtained. BPNN and PCA approaches comparatively evaluated to make the proposed model far more accurate in recognizing speakers.

5. CONCLUSION

This article examined the algorithm of speaker recognition method at different levels. The speaker recognition system is constructed using a variety of working phases, including speech filtration and speech segmentation. A description of available methods and actions is given in this article, in terms of procedure and behavior. The segmentation of the speaker is described as the separation process in silence, in content and specifying the speaker. In this article you can find the categorization of methods available for speech segmentation based on silence specific, distance specific and model. Different challenges are often explored in the context of speaker idolatry. The speech also has the problems and impurities that affect the consistency of the speech. In this article, the available methods of improving speech were discussed. This article also describes the categorization of available methods for speaking recognition. Compared with these methods, the comparative results shown that the suggested model has efficiently enhanced the recognition rate. On individual speaker recognition, the weight driven probabilistic SVM is implemented. The comparative assessment is obtained for several characters and sentence-based samples.

REFERENCES

[1] Pahar, M., & Smith, L. S. (2020, December). Coding and Decoding Speech using a Biologically Inspired Coding System. In 2020 IEEE Symposium Series on Computational Intelligence (SSCI) (pp. 3025-3032). IEEE.

[2] Yong, S., & Fuguang, Y. (2020). Application of NLP-based respiratory audio recognition framework in physical health exercise intervention. International Journal of Speech Technology, 1-13.

[3] Balaji, V. N., Srinivas, P. B., & Singh, M. K. (2021). Neuromorphic advancements architecture design and its implementations technique. Materials Today: Proceedings.

[4] Singh, M. K., Singh, A. K., & Singh, N. (2019). Multimedia analysis for disguised voice and classification efficiency. Multimedia Tools and Applications, 78(20), 29395-29411.

[5] Ghalamiosgouei, S., & Geravanchizadeh, M. (2021). Robust Speaker Identification Based on Binaural Masks. Speech Communication.

[6] Siddiqa, S. K., Apurva, K., Nandan, D., & Kumar, S. (2021). Documentation on smart home monitoring using internet of things. In ICCCE 2020 (pp. 1115-1124). Springer, Singapore.

[7] Padma, U., Jagadish, S., & Singh, M. K. (2021). Recognition of plant's leaf infection by image processing approach. Materials Today: Proceedings.

[8] Singh, M. K., Singh, A. K., & Singh, N. (2018). Disguised voice with fast and slow speech and its acoustic analysis. Int. J. Pure Appl. Math, 11(14), 241-246.

[9] Al-Hassani, R. T., Atilla, D. C., & Aydin, Ç. (2021). Development of High Accuracy Classifier for the Speaker Recognition System. Applied Bionics and Biomechanics, 2021.

[10] Vestman, V. (2020). Methods for fast, robust, and secure speaker recognition (Doctoral dissertation, Itä-Suomen yliopisto).

[11] Priya, B. J., Kunda, P., & Kumar, S. (2021). Design and Implementation of Smart Real-Time Billing, GSM, and GPS-Based Theft Monitoring and Accident Notification Systems. In Proceedings of International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications (pp. 647-661). Springer, Singapore.

[12] Singh, M. K., Singh, A. K., & Singh, N. (2019). Multimedia utilization of non-computerized disguised voice and acoustic similarity measurement. Multimedia Tools and Applications, 1-16.

[13] Sudeep, S. V. N. V. S., Venkata Kiran, S., Nandan, D., & Kumar, S. (2021). An Overview of Biometrics and Face Spoofing Detection. ICCCE 2020, 871-881.

[14] Prasanna, G. S., Pavani, K., & Singh, M. K. (2021). Spliced images detection by using Viola-Jones algorithms method. Materials Today: Proceedings.

[15] Singh, M., Nandan, D., & Kumar, S. (2019). Statistical Analysis of Lower and Raised Pitch Voice Signal and Its Efficiency Calculation. Traitement du Signal, 36(5), 455-461.

[16] Veerendra, G., Swaroop, R., Dattu, D. S., Jyothi, C. A., & Singh, M. K. (2021). Detecting plant Diseases, quantifying and classifying digital image processing techniques. Materials Today: Proceedings.

[17] Santhoshi, M. S., Sharath Babu, K., Kumar, S., & Nandan, D. (2021). An investigation on rolling element bearing fault and real-time spectrum analysis by using short-time fourier transform. In Proceedings of International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications (pp. 561-567). Springer, Singapore.

[18] Singh, M. K., Singh, A. K., & Singh, N. (2018). Acoustic comparison of electronics disguised voice using different semitones. Int. J. Eng. Technol.(UAE). https://doi. org/10.14419/ijet. v7i2, 16.

[19] Reynolds, D. A. (2002, May). An overview of automatic speaker recognition technology. In 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (Vol. 4, pp. IV-4072). IEEE.

[20] Furui, S. (1996). An overview of speaker recognition technology. Automatic speech and speaker recognition, 31-56.

International Journal of
Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 4 | Pages 862-867 | e-ISSN: 2347-470X

[21] Singh, M. K., Singh, N., & Singh, A. K. (2019, March). Speaker's Voice Characteristics and Similarity Measurement using Euclidean Distances. In 2019 International Conference on Signal Processing and Communication (ICSC) (pp. 317-322). IEEE.

[22] Kanchana, V., Nath, S., & Singh, M. K. (2021). A study of internet of things oriented smart medical systems. Materials Today: Proceedings.

[23] Furui, S. (1997). Recent advances in speaker recognition. Pattern recognition letters, 18(9), 859-872.

[24] Wang, Y. (2020). Implementation and Improvement of Common Text-Independent Speaker Identification (Doctoral dissertation, Northern Illinois University).