A REVIEW ON SPEAKER RECOGNITION: FUNDAMENTAL CONCEPTS AND CHALLENGES

SUJIYA SREEDHARAN
Ph.D Research Scholar
Department of Computer Science
Bharathiar University
Coimbatore – 641046
Sathyapriya08@yahoo.com

CHANDRA ESWARAN
Professor and Head
Department of Computer Science
Bharathiar University
Coimbatore – 641046
chandra.e@buc.edu.in

***Abstract*— In natural human-to-human interaction and communication, recognising a person by their voice is a crucial human trait that most people take for granted. Speaker recognition is a biometric system that uses particular characteristics elicited from a user's vocal utterances to authenticate their individuality. It is the process of automatically acknowledging the speaker based on the speech. Primitively, this paper briefly describes all the main aspects of Automatic Speaker Recognition, such as speaker identification, verification, diarization, feature extraction, Feature normalization, Modelling techniques and challenges in speaker recognition system.**

Keywords — *Speaker Recognition, Feature Extraction, Normalization, Modelling Techniques, Challenges in Speaker Recognition..*

I. INTRODUCTION

Voice is a behavioral biometric that communicates information about a person's characteristics, such as ethnicity, age, gender, and emotion. Speaker recognition is the process of recognizing someone's identity based on their voice [1]. Despite the fact that researchers have been working on speaker recognition for over eight decades, technological breakthroughs such as the Internet of Things (IoT), smart gadgets, voice assistants, smart homes, and humanoids have made its use fashionable these days. As numerous applications employ voice to better ordinary human life, speech signal processing technology has become a prominent communication technology. ASR is a crucial tool in digital signal processing for recognizing persons based on their voice [2,4]. As the human voice is a vital attribute of an individual, it can supply a lot of information [7,29]. The human voice contains information such as accent, language, speech, emotion, gender, and the speaker's identity. Because of variances in the forms of the vocal tract, larynx diameters, and other components of human voice production organs, each person's voice is distinct [8,9]. The pace or speed of a voice is determined by its volume, pitch level, and quality, whilst the articulation rate and speech pauses are determined by the speaker's speaking style [8]. Although multiple review papers in the field of ASR have been published, each one addresses a different aspect of the subject.

SPEECH RECOGNITION VS SPEAKER RECOGNITION

Speech recognition can help persons with a variety of disabilities, including those with physical disabilities who find typing difficult, uncomfortable, or impossible, and those with dyslexia who have trouble identifying and writing words [12][29]. Speech recognition's efficacy is significantly influenced by the language and text corpus [5], as it converts audio into text. Speaker recognition, on the other hand, is used to identify the individual who is speaking. Pitch, speaking manner, and accent are all factors that influence the disparities [11]. Speaker recognition is utilised in a variety of applications, including biometrics, security, and even human-computer interaction.

Table 1. Speaker recognition vs. Speech recognition.

Features	Speaker Recognition	Speech Recognition
Recognition	**Measures voice pattern, speaking style, and other linguistic features to identify who is speaking.**	**Recognizes what is being said and converts them into text.**
Focus	**To identify the speaker.**	**To identify and digitally record what the speaker is saying. Vocabulary of what is being said by the speaker and turns the words into digital texts.**
Application	**Voice biometrics.**	**Speech to text.**

SPEAKER RECOGNITION

Speaker recognition is a biometric system that uses particular characteristics elicited from a user's vocal utterances to authenticate their individuality. It is the automatic process of recognising the speaker based on the characteristics of the voice signal [3]. A conventional speaker recognition system assesses a person's individuality by measuring the features of their voice or speech. The most reasonable way to evolve human perceptions is through voice or speaking. ASR systems have grown in popularity with the introduction of human-computer research methods [6]. These sophisticated systems are now employed in a variety of applications, including person identification, verification, voice calling, online banking, telephone shopping, security control, and forensic applications [10]. Pre-processing, feature extraction, and speaker modelling are the three main elements of a typical speech recognition system [14]. Based on the recognition criteria, an ASR can be divided into numerous classes. The numerous types of speaker recognition approaches are shown in Figure 1. The shown recognition approaches are detailed in the following subsections.

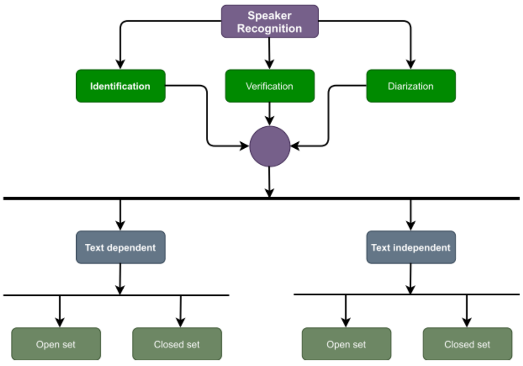

Fig. 1. Identification and verification

SPEAKER RECOGNITION CLASSIFICATION

Among the various categorization criteria, speaker recognition is classified as Automatic speaker identification (SI)[15], speaker verification (SV) [15], are often regarded as the most basic and practical methods for avoiding unauthorized access to computer systems. The following is a quick description of these subdomains of speaker recognition:

Speaker identification (SI) is a technique for determining an anonymous speaker's identity based on their spoken words. Speaker identification selects the correct speaker from a list of known voices. It is a method of locating a person using the various utterances stored in the database. This method uses a 1: N match, in which a specific utterance is compared to N templates.

Speaker verification (SV) is a technique that uses the speaker's voice to verify a specific identification. The acquired properties of the SI system are connected with all of the speakers' attributes in a voice model database. In SV systems, on the other hand, the acquired characteristics are simply related to the speaker's stated stored qualities. One speaker's speech is compared to one template in a 1:1 match. All the categories of speaker recognition work on both text dependent and text independent and in open and close-set environment.

TEXT-INDEPENDENT AND TEXT-DEPENDENT APPROACHES

Text-dependent: Text-dependent ASR approach [8]– [13] defines a method in which the test utterance is equivalent to the text used during the enrolling phase. The test speaker has prior experience with the model. The local lexicon has a low enrollment and trial stage, making it difficult to get an accurate result. Despite this, it confronts minimal scientific and technical difficulties. With feature extraction approach, speaker modelling, and score normalization using a likelihood ratio score, the first text-dependent speaker identification established the essential characteristics of the current state of the art in the 1990s [14].

Text-Independent: The voice signal's training and testing are completely unconstrained in a text-independent ASR task [15]– [19]. Both the training and testing phases of the speech signal normally take a long time to create. In this case, the test speaker has no prior knowledge of the enrollment phase samples in the testing phase. Text-independent speaker recognition, on the other hand, is more convenient than text-dependent speaker recognition since the speaker can freely communicate with the system. To improve accuracy, it will require more extensive training and testing of utterances. The open set and closed set speaker recognition challenges, which are covered in the next section.

a. OPEN-SET CLOSED-SET ENVIRONMENT

Based on the number of qualifying speakers in the system, ASR designs are classed as open set or closed set. The following are the two types of ASR methods:

- An open set method is one in which any number of trained speakers are used. Because the anonymous speech could come from a wide range of unknown speakers, this system is known as an open set [20].
- Closed set: In a closed set system, only a certain number of speakers are enrolled. The system uses this strategy to try to deduce a speaker's individuality from a collection of recorded voices [21].

II. FEATURE EXTRACTION APPROACHES IN SPEAKER RECOGNITION

In this article, utterance features are feature parameters collected from a whole utterance. Because many typical pattern-recognition algorithms operate on fixed dimension vectors, this becomes much more essential in the context of automatic speaker recognition. Acoustic/segmental features cannot be employed directly with such classifiers due to the changeable length/duration property of speech. Due to the time-varying nature and context reliance of speech, simplistic approaches such as averaging segmental features over time do not appear to be very effective in this scenario [22], [23]. Considering the rate of speech as a feature, As it is evident that two people can have the same speaking rate, this feature may not be very useful on its own. Single speaker's peculiar characteristics would be [24], [25] time-varying and context/speech sound dependent. However, high-level and long-term characteristics like as dialect, accent, and tone of voice, Prosody and speaking style/rate are both useful and beneficial when combined with low-level acoustic features[26].

a. AUDITORY VERSUS ACOUSTIC FEATURES

Auditory features are thus defined as aspects of speech that can “be heard and objectively described” by a trained listener [27]. These can be specific ways of uttering individual speech sounds (e.g., the pronunciation of the vowel sounds in the word hello can be used as auditory features). Acoustic characteristics, on the other hand, are mathematically specified parameters extracted from a voice signal by automated methods. These qualities are obviously used in automatic systems, but they're also used in computer-assisted forensic speaker recognition. Acoustic qualities include fundamental frequency (F0) and formant frequency bandwidth. Acoustic features extracted from the short-term power spectrum of speech are commonly used by automatic systems. Both auditory and acoustic characteristics have advantages and disadvantages. Two voice samples may sound extremely similar, but their acoustic characteristics are drastically different [28]. Alternatively, speech samples may sound drastically different but share acoustic characteristics [30]. As a result, it is widely understood that both auditory and acoustic characteristics are required for accurate speaker recognition [35]. It might be argued that if reverse engineering of the human auditory system [31] is effective, auditory features can be extracted using automatic methods as well. Physiological based speech features are derived by analyzing a variety of factors such as experience, personality, education, community, and communication channels. Two types of learnt based features are high-level features and prosodic & Spectro-temporal features. Phonetics, idiolect, semantics, accent, and pronunciation are all high-level characteristics [32]. Prosodic and Spectro-temporal properties, on the other hand, include pitch, energy, rhythm, duration, and temporal aspects. The non-segmental appearance of speech created in long utterances, such as prosodic features, is referred to as prosodic characteristics. Accent is a technical term that describes pitch changes, stress, rhythm, loudness, and rhythm [33].

b. SHORT-TERM VS LONG TERM FEATURES

The feature parameters can be classified as short or long term depending on their temporal span. The majority of the qualities that have been described thus far are short-term or segmental in nature. Short-term acoustic characteristics, particularly those taken from the speech spectrum, are commonly used in popular automatic systems [34]. Short-term features are also useful in auditory forensic investigation, as evidenced by direct comparisons of the "r" sound and the consonant–vowel transition [35]. Short-term parameters are frequently averaged to provide long-term characteristics (e.g., fundamental frequency, short-term spectrum). These characteristics have the advantage of being less susceptible to fluctuations caused by individual speech sounds, resulting in a more consistent measurement from a speech segment. Energy, pitch, and formant contours are other long-term properties that are measured/averaged over long time periods. Such properties have also been successfully exploited in recent automatic systems [39]–[41]. An utterance-level feature, or utterance feature for short, is a feature parameter extracted from the entirety of a voice utterance. As we move forward with the discussion of automatic systems, this concept will be quite beneficial. In Behaviour based speech features the length, dimension, and fold size of the vocal tract all influence these characteristics. Short-term spectral characteristics are physiological aspects that can be quantified from tiny spoken utterances; these characteristics are used to explain the timbre and resonance characteristics of the subpharyngeal vocal tract's short-term spectral container. Characteristics of the verbal flow are voice source properties [35].

III. FEATURE NORMALIZATION

Normalization tries to reduce the mismatch between a training and test set by adapting the distribution of scores to test conditions, for example, by shifting the means and modifying the range of variance of the score distribution. Normalization techniques at the score level are most commonly employed in speaker verification, though they can also be utilized in speaker identification, because they are quite effective at reducing the discrepancy between the claimant and the imposer [45] . Desirable properties of acoustic features (and any feature parameter in a pattern-recognition problem) is robustness to degradation. One of the desirable characteristics of an ideal feature parameter [36] is that it has this property. In actuality, it's impossible to create a feature parameter that remains completely unaltered in altered acoustic settings while also providing useful speaker-dependent data. However, employing feature-normalization approaches like cepstral mean subtraction [37], feature warping [38], relative spectra (RASTA) processing [42], and quantile-based cepstral normalization [43], these alterations can be avoided in a variety of ways.

IV. SPEAKER MODELING TECHNIQUES

The modelling technique is chosen based on the type of speech, simplicity of training, computational and storage constraints, and predicted performance [44]. Modeling Techniques are classified as follows:

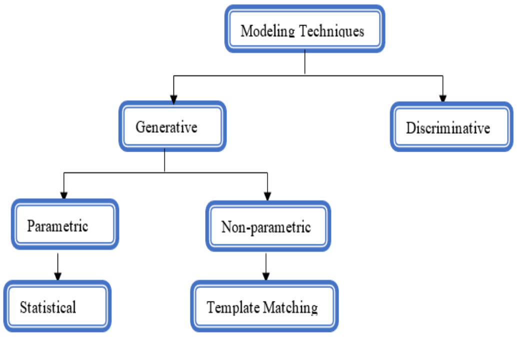

Fig. 2. Modelling techniques of Speaker Recognition

A. *Generative Model*

Generative models evaluate the distribution of features inside each speaker. They just need training data samples from the target speaker to create a statistical/non-statistical model that characterizes the feature distribution of the target speaker. Models such as the Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), and Vector Quantization are included in generative model [46].

B. Parametric Model

Parametric models presume a structure that is defined by a set of parameters. Models such as GMM and HMM are included. The density can be specified with less data because the form is constrained. The advantages of these models include efficient data utilisation, the ability to predict and interpret changes in data through changes in parameters, and the ability to use statistical summaries rather than whole data [47]. The structure is restrictive, which is a disadvantage. The GMM is a general parametric model in which each speaker is represented by a pdf (probability density function) that governs the distribution of his or her feature vectors. It has the potential to produce irregularly shaped smooth densities [11][48]. The number of training samples necessary for successful density estimation in GMM grows exponentially with the number of features. This is also known as the Dimensionality Curse [12]. GMM can also make advantage of temporal information, such as state transition probabilities, to produce continuous HMM.

C. Non-Parametric Model

In non-parametric models, just the most basic assumptions about pdf are made. Nearest Neighbour (NN)[23], Vector Quantization (VQ) [24], and Dynamic Time Warping (DTW)[49] are the most used approaches. The centroid of each feature cluster is calculated in VQ. A codebook is a collection of centroids. The Linde Buzo Gray (LBG) algorithm is used to produce a codebook for each speaker [50]. The identified speaker has the least amount of distortion. The only difference between VQ and the NN model is that distances between nearest data representations are measured. As a result, it requires less processing and memory than NN. VQ has the advantage of requiring less storage for spectral analysis data. We can attach a phonetic label with each codebook due to the distinct representation of speech sounds. Because there are finite numbers of codebook vectors, VQ has inherent spectrum distortion. As a result, there is a certain level of quantization error. Quantization error diminishes as the size of the codebook grows [50].

D. Discriminative Modelling

Models that discriminate between speakers are called discriminative models. These require training data for both target and non-target speakers in order to determine the best separation between them. Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs) are among the models included (SVMs). Soft computer modelling is another name for these modelling techniques. Rather than developing individual speaker models, the decision function between speakers is trained [28]. Flexible architectures and discriminating training power are advantages of these models, however the best structure is determined by trial and error. SVM classifiers use a non-linear boundary to separate complex regions. With far less training data, SVMs can achieve comparable or better performance than GMM.

V.CHALLENGES IN SPEAKER RECOGNITION

Limited data and lexicon [25]: For modern industrial purposes, training periods are typically made up of repetitive duplications of the enrollment lexicon. The trial period is based on a one-of-a-kind replication of a subset of the recorded lexicon over the course of 4-5 seconds of speech input. Studies show that end consumers prefer shorter registration and testing sessions, therefore some obligations are limited. In the most serious cases, the testing lexicon is selected because it closely matches the enrollment lexicon.

Channel Usage [12] : Top clients in actual applications use a variety of phones, including landlines, payphones, cordless phones, cell phones, and so on. This increases the impact of their influence on channel utilisation efficiency. A cross-channel endeavour is defined as a measurement interval that begins with a different channel from the one used throughout the training period. It's an important area where speech recognition architectures need to be improved.

Speaker model ageing [45]: There are several sources of speaker model ageing, including natural ageing, channel usage, and behavioural changes. The physiological changes that occur to the phonatory device over time are associated with biological ageing. Changes in channel usage over time may allow the speaker model to become archaic in terms of current channel usage. As a result, users' behaviour changes as they become more vulnerable to the speech interface and restructure how they interact. In conclusion, these factors have an impact on the models and scores, which is represented in the efficiency.

VI. CONCLUSION

Speaker recognition is an eminent research domain widely investigated and integrated into numerous systems to identify or verify individuals. Speaker recognition is an eminent research domain widely investigated and integrated into numerous systems to identify or verify individuals. Speaker recognition has been studied actively for several decades. The article explored the fundamentals of speaker recognition with its categories, and discussed about various feature extraction techniques conveyed via the speech spectrum, and discussed about various modelling techniques based on parametric and non-parametric division model. This article has given an extensive study on the challenges persist on speaker recognition system.

References

[1] Z. Bai and X.-L. Zhang, ‘‘Speaker recognition based on deep learning: An overview,’’ Neural Netw., vol. 140, pp. 65–99, Aug. 2021.

[2] Sujiya.S, Dr.E.Chandra “A review on Speaker Recognition” ,International Journal of Engineering and Technology, Volume 09, Year 2017, Pages 1592- 1598.

[3] R. V. Pawar, R. M. Jalnekar, and J. S. Chitode, ‘‘Review of various stages in speaker recognition system, performance measures and recognition toolkits,’’ Anal. Integr. Circuits Signal Process., vol. 94, no. 2, pp. 247–257, Feb. 2018.

[4] J. A. Gómez-García, L. Moro-Velázquez, and J. I. Godino-Llorente, ‘‘On the design of automatic voice condition analysis systems. Part II: Review of speaker recognition techniques and study on the effects of different variability factors,’’ Biomed. Signal Process. Control, vol. 48, pp. 128–143, Feb. 2019.

[5] M. Hamidi, H. Satori, N. Laaidi, and K. Satori, ‘‘Conception of speaker recognition methods: A review,’’ in Proc. 1st Int. Conf. Innov. Res. Appl. Sci., Eng. Technol. (IRASET), Apr. 2020, pp. 1–6.

[6] D. Sztahó, G. Szaszák, and A. Beke, ‘‘Deep learning methods in speaker recognition: A review,’’ 2019, arXiv:1911.06615.

[7] A. Irum and A. Salman, ‘‘Speaker verification using deep neural networks: A review,’’ Int. J. Mach. Learn. Comput., vol. 9, no. 1, pp. 1–6, 2019.

[8] V. Vestman, D. Gowda, M. Sahidullah, P. Alku, and T. Kinnunen, ‘‘Speaker recognition from whispered speech: A tutorial survey and an application of time-varying linear prediction,’’ Speech Commun., vol. 99, pp. 62–79, May 2018.

[9] A. P. Singh, R. Nath, and S. Kumar, ‘‘A survey: Speech recognition approaches and techniques,’’ in Proc. 5th IEEE Uttar Pradesh Sect. Int. Conf. Electr., Electron. Comput. Eng. (UPCON), Nov. 2018.

[10] M. Farrús, ‘‘Voice disguise in automatic speaker recognition,’’ ACM Comput. Surv., vol. 51, no. 4, pp. 1–22, Sep. 2018.

[11] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, ‘‘LibriTTS: A corpus derived from LibriSpeech for text-to-speech,’’ 2019.

[12] C. Lüscher, E. Beck, K. Irie, M. Kitza, W. Michel, A. Zeyer, R. Schlüter, and H. Ney, ‘‘RWTH ASR systems for LibriSpeech: Hybrid vs attention—W/O data augmentation,’’ 2019, arXiv:1905.03072.

[13] J. A. C. Nunes, D. Macêdo, and C. Zanchettin, ‘‘AM-MobileNet1D: A portable model for speaker recognition,’’ in Proc. Int. Joint Conf. Neural Netw. (IJCNN), Jul. 2020, pp. 1–8.

[14] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, ‘‘Speaker recognition for multi-speaker conversations using X-vectors,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2019, pp. 5796–5800

[15] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, L. P. García-Perera, F. Richardson, R. Dehak, P. A. Torres-Carrasquillo, and N. Dehak, ‘‘State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and speakers in the wild evaluations,’’ Comput. Speech Lang., vol. 60, Mar. 2020, Art. no. 101026.

[16] P. Dhakal, P. Damacharla, A. Javaid, and V. Devabhaktuni, ‘‘A near real-time automatic speaker recognition architecture for voice-based user interface,’’ Mach. Learn. Knowl. Extraction, vol. 1, no. 1, pp. 504–520, Mar. 2019.

[17] R. Jahangir, Y. W. Teh, N. A. Memon, G. Mujtaba, M. Zareei, U. Ishtiaq, M. Z. Akhtar, and I. Ali, ‘‘Text-independent speaker identification through feature fusion and deep neural network,’’ IEEE Access, vol. 8, pp. 32187–32202, 2020.

[18] X. Qin, H. Bu, and M. Li, ‘‘HI-MIA: A far-field text-dependent speaker verification database and the baselines,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2020, pp. 7609–7613.

[19] D. Zhou, L. Wang, K. A. Lee, M. Liu, and J. Dang, ‘‘Deep discriminative embedding with ranked weight for speaker verification,’’ in Proc. Int. Conf. Neural Inf. Process. Cham, Switzerland: Springer, 2020.

[20] Y. Fan, J. W. Kang, L. T. Li, K. C. Li, H. L. Chen, S. T. Cheng, P. Y. Zhang, Z. Y. Zhou, Y. Q. Cai, and D. Wang, ‘‘CN-CELEB: A challenging Chinese speaker recognition dataset,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2020, pp. 7604–7608

[21] Sujiya Sreedharan, Dr.E.Chandra “Speech Feature Encryption using Rc4 Key based AES for Speaker Verification” Wulfenia Journal, Vol.No:27 page no:19-36, 2020

[22] D. Snyder, J. Villalba, N. Chen, D. Povey, G. Sell, N. Dehak, and S. Khudanpur, ‘‘The JHU speaker recognition system for the VOICES 2019 challenge,’’ in Proc. INTERSPEECH, 2019, pp. 2468–2472.

[23] S.Poovarasan, Dr.E.Chandra “Speech Enhancement Using Sliding Window Empirical Mode Decomposition and Hurst-based Technique” Archives of Acoustics, Volume 44, Year 2019, Pages 429–437.

[24] Y. Zhang, H. Yu, and Z. Ma, ‘‘Speaker verification system based on deformable CNN and time-frequency attention,’’ in Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC), 2020, pp. 1689–1692.

[25] H. Zeinali, S. Wang, A. Silnova, P. Matějka, and O. Plchot, ‘‘BUT system description to VoxCeleb speaker recognition challenge 2019,’’ 2019, arXiv:1910.12592. [Online]. Available: http://arxiv.org/abs/1910.12592

[26] J. S. Chung, J. Huh, S. Mun, M. Lee, H. S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, ‘‘In defence of metric learning for speaker recognition,’’ 2020, arXiv:2003.11982.

[27] C. Zhang, F. Bahmaninezhad, S. Ranjan, H. Dubey, W. Xia, and J. H. L. Hansen, ‘‘UTD-CRSS systems for 2018 NIST speaker recognition evaluation,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2019, pp. 5776–5780.

[28] J.-W. Jung, H.-S. Heo, J.-H. Kim, H.-J. Shim, and H.-J. Yu, ‘‘RawNet: Advanced end-to-end deep neural network using raw waveforms for textindependent speaker verification,’’ 2019, arXiv:1904.08104.

[29] S.Poovarasan, Dr.E.Chandra “Comparative Analysis of Various Noise Types using Empirical Mode Decomposition Based Hurst Exponent Techniques” International Journal of Scientific & Technology Research, Volume 08, Year 2019, Pages 1693- 1696.

[30] S. Madikeri, P. Motlicek, and S. Dey, ‘‘A Bayesian approach to inter-task fusion for speaker recognition,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2019, pp. 5786–5790.

[31] R. Li, J.-Y. Jiang, J. L. Li, C.-C. Hsieh, and W. Wang, ‘‘Automatic speaker recognition with limited data,’’ in Proc. 13th Int. Conf. Web Search Data Mining, Jan. 2020, pp. 340–348.

[32] Sujiya Sreedharan, Dr.E.Chandra “Speech Encryption using Advanced Encryption

[33] Standard for Secured Communication” International Journal of Recent Technology and Engineering, ISSN: 2277-3878, Volume-8 Issue-3, September 2019

[34] S. M. Kye, Y. Jung, H. B. Lee, S. J. Hwang, and H. Kim, ‘‘Meta-learning for short utterance speaker recognition with imbalance length pairs,’’ 2020, arXiv:2004.02863. [Online].

[35] A. Chowdhury and A. Ross, ‘‘Fusing MFCC and LPC features using 1D triplet CNN for speaker recognition in severely degraded audio signals,’’ IEEE Trans. Inf. Forensics Security, vol. 15, pp. 1616–1629, 2020.

[36] K. Saakshara, K. Pranathi, R. M. Gomathi, A. Sivasangari, P. Ajitha, and T. Anandhi, ‘‘Speaker recognition system using Gaussian mixture model,’’ in Proc. Int. Conf. Commun. Signal Process. (ICCSP), Jul. 2020, pp. 1041–1044.

[37] S. Biswas and S. S. Solanki, ‘‘Speaker recognition: An enhanced approach to identify singer voice using neural network,’’ Int. J. Speech Technol., vol. 24, pp. 1–13, Mar. 2020.

[38] S.Sujiya,Dr.E.Chandra “A Review on Speaker Verification: Challenges and Issues”, International Journal of Scientific & Technology Research, Volume 08, Year 2019, Pages 956-960.

[39] S. A. El-Moneim, M. Nassar, M. I. Dessouky, N. A. Ismail, A. S. El-Fishawy, and F. E. A. El-Samie, ‘‘Text-independent speaker recognition using LSTM-RNN and speech enhancement,’’ Multimedia Tools Appl., vol. 79, no. 33, pp. 24013–24028, 2020.

[40] M. Tripathi, D. Singh, and S. Susan, ‘‘Speaker recognition using SincNet and X-vector fusion,’’ in Proc. Int. Conf. Artif. Intell. Soft Comput. Cham, Switzerland: Springer, 2020, pp. 252–260.

[41] N. Tawara, A. Ogawa, T. Iwata, M. Delcroix, and T. Ogawa, ‘‘Framelevel phoneme-invariant speaker embedding for text-independent speaker recognition on extremely short utterances,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2020, pp. 6799–6803.

[42] R. Peri, M. Pal, A. Jati, K. Somandepalli, and S. Narayanan, ‘‘Robust speaker recognition using unsupervised adversarial invariance,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2020, pp. 6614–6618.

[43] L. Lu, L. Liu, M. J. Hussain, and Y. Liu, ‘‘I sense you by breath: Speaker recognition via breath biometrics,’’ IEEE Trans. Dependable Secure Comput., vol. 17, no. 2, pp. 306–319, Mar. 2020’

[44] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe, ‘‘End-to-end neural speaker diarization with permutationfree objectives,’’ 2019, arXiv:1909.05952.

[45] A. Zhang, Q. Wang, Z. Zhu, J. Paisley, and C. Wang, ‘‘Fully supervised speaker diarization,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2019, pp. 6301–6305.

[46] Karthika Kuppusamy, Chandra Eswaran ” Convolution and Deep Neural Networks based techniques for extracting the age - relevant features of the speaker ” , Journal of Ambient Intelligence and Humanized Computing, Year 2021, Pages: 1-13.

[47] Poovarasan, S., & Chandra, E “A Variant of SWEMDH Technique based on Variational Mode Decomposition for Speech Enhancement”, International Journal of Knowledge-Based and Intelligent Engineering Systems, Volume 25, Issue 3, Pages 299-308, Year 2021.

[48] Sujiya Sreedharan, Dr.E.Chandra “A light weight encryption scheme using Chebyshev polynomial maps”, Optik, Volume 240, Year 2019.

[49] K.Karthika, Dr.E.Chandra “Speech and Speaker Recognition: A Review”, International Journal of Scientific & Technology Research, Volume 08, Year 2019, Pages 938- 944.

[50] Sujiya.S,Dr.E.Chandra “A Comparative Analysis of Feature Extraction Techniques for Speaker Verification” Interciencia Journal, Volume 43, Year 2018, Pages 245-264.

[51] I. Bisio, C. Garibotto, A. Grattarola, F. Lavagetto, and A. Sciarrone, ‘‘Smart and robust speaker recognition for context-aware in-vehicle application” IEEE Trans. Veh. Technolo., vol. 67, no. 9, Sep. 2018.