Voice-Activated Biometric Systems

Fidalizia Pyrtuh

Department of Electronics and Communication Engineering

North-Eastern Hill University

Shillong, India

fpyrtuh@nehu.ac.in

**ABSTRACT**

As technology advances at an unprecedented rate, biometric systems have emerged as a revolutionary approach to authentication and identification. Leveraging unique physiological and behavioral characteristics, these systems offer robust security and seamless user experiences across various applications. This paper explores the future trends in biometric systems, focusing on the innovations poised to transform the landscape of digital security and personal identification using voice as a password. Multi-modal biometrics, continuous authentication, and privacy-preserving techniques are expected to enhance accuracy, thwart spoofing attempts, and safeguard user data. Additionally, the integration of biometrics with artificial intelligence (AI) promises to unlock new levels of efficiency and adaptability. The paper delves into the ethical considerations surrounding the widespread adoption of biometrics and addresses the challenges that lie ahead.

**Keywords**—biometric; voice recognition;

* **INTRODUCTION**

Biometric solutions have emerged as a game changer in meeting the ever-increasing demand for safe and seamless identification and authentication. Biometrics provides a robust and reliable technique of recognizing an individual, allowing them to engage with their digital systems, access sensitive data, and perform transactions with heightened security by using their unique physiological and behavioral attributes. Behavioral biometrics include voice, signatures, keystrokes, typing, and other behavioral biometrics, whereas physiological biometrics include iris, face, fingerprints, retina, ear, and DNA. Voice biometrics has become a popular tool to authenticate customers who receive help through interactive voice response (IVR) phone systems. Voice recognition provides a quick, automated way to obtain customer account information and confirm the caller's identification. Because of advancements in artificial intelligence and machine learning, biometric data can now be promptly processed, making the technology more beneficial and approachable. The need for biometrics as a technique of multi-factor authentication has increased due to the rise of terrorism and cybercrime.

It is a common experience to identify familiar people by their voices alone like friends, colleagues, family members, etc. The ability of human beings to identify speakers by their voice or speech sounds is known as Speaker recognition. Each speaker has a different style of speech delivery, and vocabulary usage, because of the unique physiological structure of their speech production system. This structure includes distinctive size and shape of the vocal tract, mouth, nasal cavity, and larynx[1,2]. It also contains information related to the behavioral aspects of a speaker such as an accent, involuntary transforms of acoustic parameters, etc. This causes differences between the speakers in the speech production process. The rapid growth of biometrics-based technology, measuring human behavioral or physiology characteristics for identifying and verifying an individual leads to technological curiosity on the mechanics of realization of human speech to ease/automate Human-Machine Interaction, which in turn leads to extensive research on the automatic speaker/ speech recognition process.

The implementation of face recognition as a contactless method of identity verification has been encouraged during the COVID pandemic.. As technology continues to evolve, Voice recognition biometric systems hold immense promise, heralding an era of advanced capabilities and transformative applications. This paper delves into the future trends that are set to shape the landscape of biometric systems in the coming years. We explore innovative approaches that aim to enhance accuracy, convenience, and privacy in the domain of digital security.

The integration of biometrics with Artificial Intelligence (AI) and machine learning, holds the key to unlocking new frontiers in personalized user experiences. While the future of biometric systems appears promising, it is essential to navigate the ethical and privacy challenges that arise from their widespread adoption. As these systems collect and process sensitive biometric data, concerns regarding data protection, consent, and potential biases warrant careful consideration. Throughout this paper, we aim to provide an insightful exploration of the exciting future trends in Voice recognition biometric systems, covering their potential applications in various industries, the impact of AI integration, and the ethical considerations surrounding their deployment. By examining these emerging developments, we seek to shed light on the transformative potential of biometrics in reshaping the digital security landscape and creating a safer, more user-centric world.

* **FUNDAMENTALS OF VOICE AS A BIOMETRIC**

The speech signal communicates many levels of information. At the primary level, it conveys the information via words. At other levels, the message contains information like language spoken, emotion, gender, and in general, the identity of a speaker. The main aim of an automatic speaker recognition system is to extract, characterize and recognize the message/ information in the speech signal, hence conveying the speaker's identity. Speaker Recognition is the process of automatically identifying or extracting a person’s personal identity by analyzing an individual spoken utterance [3-6]. Speaker Recognition activity can be divided into two principal tasks, Speaker Identification(SI) and Speaker Verification(SV). The main difference between identification and verification is in identification, the number of decisions is equal to the size of the population, whereas, in verification, there is acceptance or rejection, irrespective of the population size. Speaker recognition is considered as a general organic process whereas Speaker Identification and Verification refers to the tasks associated with the Speaker Recognition process. Speaker Identification relates to the task of determining a speaker’s identity whereas Speaker Verification refers to the task of validating a speaker’s claimed identity[7-9].

In the case of forensic application of SV systems, it is less likely to get sufficient data for the enrolment process[10]. In access control-type cases, the average utterance length is restricted to a few seconds only[11]. Hence, it is important to take up research efforts to get reliable SV performance in short-duration conditions. Over the years, the effort to develop an SV system suitable for real-world implementation has experienced significant progress[12-14]. Rigorous reviews on SV systems are presented in[1,2,15], considered overall issues and techniques in SV, and also mentioned limited duration as one of the problems in ASV[16].

The area of SV has witnessed a major breakthrough in past decades with the development of various techniques suitable for modeling speaker characteristics. These advancements in the field of SV have opened doors toward practical deployable systems for person authentication. There have been several attempts made for the development of person authentication systems in practical scenarios like Remote person authentication[17], smart home security applications [18], speech biometric attendance systems [19], verification over a telephone-based network[20] showcase SV with short utterances, as an emerging area for having deployable systems with real-world applications.

The short utterance-based scenario comes into the picture when the focus is on person authentication for practical deployment as the time involved in authenticating a person is very less with most of the other biometric attribute-based systems used in practice. The use of short utterances for SV not only reduces the required testing time but also provides comfort to the speakers as they are less burdened with speaking. In this regard, the text-dependent SV is found as a favorable candidate for having a deployable system due to less amount of train/test time involved. However, there is a very high chance of spoofing by unauthorized users as the fixed phrase is global across all the users of the system.

* **VOICE RECOGNITION SYSTEMS**

The fundamentals of voice recognition as a biometric technology involve the following key aspects:

1. Database collection:

Each user's voiceprint is created as the first stage in voice recognition. A voiceprint is a distinctive illustration of a person's vocal traits that includes information on pitch, tone, accent, cadence, and pronunciation. When a person enrolls, their voices are often captured, analyzed, and used to produce a template that will be saved in the system's database.

2. Feature Extraction:

The system takes the voiceprint it has made and extracts its key characteristics. These characteristics are chosen based on how well they can reliably identify one person from another. Mel-frequency cepstral coefficients (MFCCs), LPCC, and x-vectors, which effectively describe vocal features, are frequently utilized in speech recognition systems.

3. Pattern Matching:

The speech recognition system records the user's voice sample and extracts its attributes during the verification or identification procedure. Using a pattern-matching process, The retrieved features are compared with the database's stored voiceprints. The technology verifies the speaker's identification if the features closely resemble a voiceprint that is already saved.

4. Validation:

A threshold is set for validating the test and training data accuracy.

The major advancements in SV research are due to improvements in the classifier domain. The primitive SV systems were developed using vector quantization (VQ)[21], a dynamic time warping[22-25] approach. Later on, with the introduction of GMM[26], ASV research has evolved in the past two decades with more focus on channel compensation[27], data variability, etc. GMM with the UBM[28] was proposed with significant improvement over independently trained GMM using the maximum-likelihood(ML) approach[29]. The latent variable approach has introduced another new paradigm in ASV technology. For example, factor analysis(FA)-based approaches were proposed to model the intersession variability in the context of GMM supervector[30]. Motivated by the success of joint FA(JFA), that is, speaker factors directly as features for classification, Dehak et al.[31] introduced single total-variability(TV) subspace-based modeling of the speakers, unlike separate subspaces for speakers and channels in JFA. Recent SV technology focused on TV modeling, also known as i-vector. The i-vector space is further modeled using a separate speaker and channel-dependent subspaces with Gaussian probabilistic LDA(GPLDA)[32], and this approach efficiently handles the intersession variability. With the help of the current ASV technology's i-vector method, variable-length speech utterances can be efficiently represented in fixed dimensions. In recent times, deep-learning-based approaches have attracted much attention and caused extensive interest in various domains[33]. The studies used Deep Neural Network (DNN) models trained for voice recognition to create UBM for SV, such as acoustic models, with rich data phones that may be used to create background models that are more effective[34–38]. DNNs are successfully used to extract speaker information features.

Voice recognition is becoming a more widely used and trustworthy biometric tool for identifying people based on their distinctive speech characteristics. For purposes of identification and authentication, biometric systems often rely on the distinctive physical and behavioral characteristics of individuals. Voice recognition, commonly referred to as speaker recognition, is a type of behavioral biometrics that examines speech patterns to identify behavioral patterns. The fundamentals of voice recognition as a biometric technology involve the following key aspects:

1. Database creation:

Each user's voiceprint is created as the first stage in voice recognition. A voiceprint is a distinctive illustration of a person's vocal traits that includes information on pitch, tone, accent, cadence, and pronunciation. When a person enrolls, their voices are often captured, analyzed, and used to produce a template that will be saved in the system's database.

2. Feature Extraction:

The system takes the voiceprint it has made and extracts its key characteristics. These characteristics are chosen based on how well they can reliably identify one person from another. Mel-frequency cepstral coefficients (MFCCs), which effectively describe vocal features, are frequently utilized in speech recognition systems.

3. Pattern Matching:

The speech recognition system records the user's voice sample and extracts its attributes during the verification or identification procedure. Using a pattern-matching process, The retrieved features are compared with the database's stored voiceprints. The technology verifies the speaker's identification if the features closely resemble a voiceprint that is already saved.

4. Text-Dependent vs. Text-Independent Systems:

There are two primary categories of voice recognition systems: text-dependent and text-independent. Users are required to say a certain phrase or a group of phrases throughout the authentication procedure in text-dependent systems. Though less user-friendly, these technologies are more secure. On the other hand, text-free systems let users speak freely, making them more user-friendly but may be less secure.

* **FUTURE TRENDS IN VOICE RECOGNITION**

The research effort to tackle the problem of short utterances for SV systems has significantly increased in recent years[17, 39]. The issue is dealt with at many levels of the ASV framework's sub-systems, including feature extraction, speaker modeling, score normalization, and score calibration. The conventional state-of-art speaker recognition models perform poorly with short utterances. Hence, a study of different pattern recognition approaches and normalization techniques, that are capable of handling limited data short utterances, can be a window of exploration for identifying a better recognition system. This research work shall relate to finding/exploring various procedures and algorithms to handle limited data on short utterance-based SV.

The accuracy and dependability of the personal authentication system may be improved by combining different biometrics.

- Security and access control: Because each person's voice is unique and difficult to reproduce or fabricate, speaker recognition provides a strong and secure authentication technique. Real-time authentication is provided using speaker recognition, allowing for quick and seamless access control. Speaker recognition, unlike some biometric technologies such as fingerprint scanning, does not involve physical touch, making it a non-intrusive form of identification. Speaker recognition can be used for remote verification over the phone or through voice-controlled applications, increasing the authentication process's versatility.

- Forensic Speaker Recognition: speech biometrics are employed in law enforcement to match speech samples from crime scenes, such as threatening phone calls, ransom requests, or recorded messages left by suspects. Investigators can restrict the list of suspects and obtain important information by comparing these speech samples to known voiceprints of prospective suspects or individuals of interest. Speaker recognition enables voice lineups, which are similar to traditional police lineups in which eyewitnesses identify suspects from a lineup of images. Witnesses or victims who have heard the perpetrator's voice can listen to voice samples of several suspects to choose which one they believe matches the perpetrator's voice. Voice biometrics can help establish if the voice in a recorded message matches the person making the threat, blackmail, or extortion.

- Voice Biometrics in Commercial Products: - Commercial Applications of Voice Biometrics: Virtual voice assistants, such as Amazon Alexa, Apple's Siri, Google Assistant, or Microsoft's Cortana, are frequently found in voice-controlled gadgets. These voice assistants use speaker recognition to identify users and tailor responses based on their preferences and previous interactions. Individuals with hearing difficulties may be able to use speaker recognition because it focuses on speech qualities rather than listening abilities.

Over the decades, voice biometrics has had a wide range of applications eg. customize services or information by voice, Intelligence applications, finance, banking and surveillance, and criminal-forensic investigations. Despite the advancements in speaker recognition technology, challenges remain, such as dealing with background noise, changes in a user's voice due to factors like illness, and attempts to spoof or deceive the system. Ongoing research focuses on enhancing accuracy, robustness, and adaptability to different speaking conditions.

* **CONCLUSION**

Voice recognition offers advantages and limitations as a biometric technology. Among the obstacles are variations in voice induced by aging, health issues, or emotional events. Ambient noise and other factors may also have an impact on the accuracy of voice recognition systems. However, voice recognition has a number of advantages. It is a non-invasive biometric approach because it just uses the user's speech for authentication. Users can authenticate their identities easily and conveniently by speaking, making it user-friendly and convenient. Voice biometrics can also be used for forensic investigations, telephone-based authentication, and access control, among other things.

**REFERENCES**

1). J. Campbell, “Speaker recognition: A tutorial,” Proc. IEEE, vol. 85, pp. 1437–1462, 9 1997.  
2). T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: from features to supervectors,” Speech Commun., vol. 52, pp. 12–40, 1 2010  
3). G. R. Doddington, “Speaker recognition-identifying people by their voices,” vol. 73, 1985, pp. 1651–1664.  
4). F. Nolan, The phonetic basis of speaker recognition. Cambridge studies in speech science and communication. Cambridge: Cambridge University Press, 1983.  
5). K. R. S. A. Schmidt, “Identification of known voices as a function of familiarity and narrow-band coding,” Journal of Acoustic Society of America, vol. 77, pp. 658–663, 1985.  
6). K. E. D. Van Lacker J. Kreiman, “Familiar voice recognition:patterns and parameters part i: Recognition of backward voices,” Journal of Phonetics, vol. 13,pp. 19–38, 1985.

7). S. Chakroborty and G. Saha, “Improved text-independent speaker identification using fused mfcc & imfcc feature sets based on gaussian filter,” Int. J. Signal Process, vol. 5, pp. 11–19, 1 2009.  
8). J. Campbell, “Speaker recognition: A tutorial,” Proceedings of the IEEE, vol. 85, no. 9, pp. 1437–1462, 1997.  
9). J. Naik, “Speaker verification: A tutorial,” IEEE Communications Magazine, vol. 28, no. 1, pp. 42–48, 1990.

10). M. I. Mandasari, M. McLaren, and D. Van Leeuwen, “Evaluation of i-vector speaker recognition systems for forensic application.,” Jan. 2011, pp. 21–24.  
11). A. Larcher, K. A. Lee, B. Ma, and H. Li, “The rsr2015: Database for text-dependent speaker verification using multiple pass-phrases,” 13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012, vol. 2, pp. 1578–1581, Jan. 2012.

12). J. Colombi, D. Ruck, T. Anderson, S. Rogers, and M. Oxley, “Cohort selection and word grammar effects for speaker recognition,” in 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, vol. 1, 1996, 85–88 vol. 1.  
13). S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4, pp. 357–366, 1980.  
14). A. L. Higgins, L. G. Bahler, and J. E. Porter, “Speaker verification using randomized phrase prompting,” Digit. Signal Process., vol. 1, pp. 89–106, 1991.

15). H. S. Jayanna and S. R. M. Prasanna, “Analysis, feature extraction, modeling and testing techniques for speaker recognition,” IETE Technical Review, vol. 26, no. 3, pp. 181–190, 2009.  
16). R. Teunen, B. Shahshahani, and L. Heck, “A model-based transformational approach to robust speaker recognition,” in INTERSPEECH, 2000.

17). D. Chakrabarty, S. R. Prasanna, and R. K. Das, “Development and evaluation of online text-independent speaker verification system for remote person authentication,” vol. 16, no. 1, 2013.  
18). K. A. Lee, A. Larcher, H. Thai, B. Ma, and H. Li, “Joint application of speech and speaker recognition for automation and security in smart home.,” Jan. 2011, pp. 3317–3318.  
19). S. Dey et al., “Speech biometric based attendance system,” Feb. 2014, pp. 1–6.  
20). R. K. Das, S. Jelil, and S. R. Mahadeva Prasanna, “Development of multi-level speech based person authentication system,” vol. 88, no. 3, 2017

21). S. Furui, “Cepstral analysis technique for automatic speaker verification,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 29, no. 2, pp. 254–272, 1981.

22). G. Al-Naymat, S. Chawla, and J. Taheri, “Sparsedtw: A novel approach to speed up dynamic time warping,” Conferences in Research and Practice in Information Technology Series, vol. 101, Jan. 2012.

23). S. Salvador and P. K.-F. Chan, “Fastdtw: Toward accurate dynamic time warping in linear time and space,” 2004.

24). T. Zaharia, S. Segarceanu, M. Cotescu, and A. Spataru, “Quantized dynamic time warping (dtw) algorithm,” in 2010 8th International Conference on Communications, 2010, pp. 91–94.

25). D. A. Reynolds, “Speaker identification and verification using gaussian mixture speaker models,” Speech Communication, vol. 17, no. 1, pp. 91–108, 1995.

26). A. Solomonoff, W. Campbell, and I. Boardman, “Advances in channel compensation for svm speaker recognition,” in Proceedings. (ICASSP ’05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005., vol. 1, 2005, I/629–I/632 Vol. 1. “Speaker verification using adapted gaussian mixture models,” Digital Signal Processing, vol. 10, no. 1, pp. 19–41, 2000.

27). P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, “A study of inter-speaker variability in speaker verification,” Audio, Speech, and Language Pro-cessing, IEEE Transactions on, vol. 16, pp. 980–988, Aug. 2008.

28). N. Dehak, R. Dehak, P. Kenny, and P. Dumouchel, “Comparison between factor analysis and GMM support vector machines for speaker verification,” in Proceedings of the Speaker and Language Recognition Workshop (IEEE-Odyssey 2008), Stellenbosch, South Africa, 2008.

29). P. Kenny, “Bayesian speaker verification with heavy-tailed priors,” in Odyssey, 2010.

30). L. Deng and D. Yu, “Deep learning: Methods and applications,” vol. 7, no. 3–4, 2014.

31). P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and M. J. Alam, “Deep neural networks for extracting baum-welch statistics for speaker recognition,” Jun. 2014.

32). Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, “A novel scheme for speaker recognition using a phonetically-aware deep neural network,” in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014,pp. 1695–1699.

33). S. Sremath Tirumala and S. R. Shahamiri, “A review on deep learning approaches in speaker identification,” Nov. 2016, pp. 142–147.

34). C. Yang, W. Yang, and S. Wang, “Based on artificial neural networks for voice recognition word segment,” in 2011 IEEE 3rd International Conference on Communication Software and Networks, 2011, pp. 394–396.

35). D. Ying, Y. Yan, J. Dang, and F. K. Soong, “Voice activity detection based on an unsupervised learning framework,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 8, pp. 2624–2633, 2011.

36). L. Li, D. Wang, Z. Zhang, and F. Zheng, “Deep speaker vectors for semi-text-independent speaker verification,” May 2015.

37). P. Melin, J. Ur ́ıas, D. Solano, M. Soto, M. A. Lopez, and O. Castillo, “Voice recognition with neural networks, type2 fuzzy logic and genetic algorithms,” Engineering Letters, vol. 13, pp. 108–116, Aug. 2006.

38). E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, “Deep neural networks for small footprint text-dependent speaker verification,” in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 4052–4056.

39). A. K. Vuppala and K. S. Rao, “Vowel onset point detection for noisy speech using spectral energy at formant frequencies,” IEEE Trans. Audio, Speech, LanguageProcessing, vol. 17, pages 229–235, 2009