Voice-Activated Biometric Systems

Fidalizia Pyrtuh

Department of Electronics and Communication Engineering

North-Eastern Hill University

Shillong, India

fpyrtuh@nehu.ac.in

**ABSTRACT**

With the progress of technology at an unparallel rate, biometric systems have developed to be a revolutionary method for authentication and identification. These systems provide a reliable level of security and an efficient user experiences across various applications by exploiting distinct physiological and behavioral characteristics. This paper examines the future trends in biometric systems, targeting the innovations poised for altering the landscape of digital security and personal identification utilizing voice as a password. Multi-modal biometrics, continuous authentication, and privacy-preserving techniques are anticipated to improve accuracy, thwart spoofing attempts, and secure user data. Moreover, unlocking new levels of efficiency and adaptability is assured by merging biometrics with artificial intelligence (AI). The paper investigates the ethical considerations encompassing the extensive adoption of biometrics and cope with the upcoming challenges.

* **INTRODUCTION**

Biometric solutions have developed to be game changer in satisfying the growing need for safe and seamless identification and authentication. By utilizing their distinct physiological and behavioral characteristics, Biometrics offer a robust and authentic technique of identifying an individual, enabling them to be associated with their digital systems, gaining access to confidential data, and carry out transactions with maximum security. Voice, signatures, keystrokes and typing constitute the behavioral biometrics whereas physiological biometrics comprises iris, face, fingerprints, retina, ear, and DNA. Voice biometrics has emerged a prominent tool for customer authentication assisted by interactive voice response (IVR) phone systems. Voice recognition eases access to customer account information and confirming the caller's identity in a faster and automated way. The advancements in artificial intelligence and machine learning help in handling biometric data promptly, thereby the technology is made more advantageous and approachable. Rise of terrorism and cybercrime in today’s world has led to the growth in demand for biometrics as a tool for multi-factor authentication.

Identifying familiar persons such as friends, colleagues and family members by their voices alone is a general experience. Speaker recognition is referred to as the capability of human beings in identifying speakers by their voice or speech sounds. The distinct physiological structure of speech production system in each speaker enables them to possess dissimilar style of speech delivery and vocabulary usage. This structure includes distinctive size and shape of the vocal tract, mouth, nasal cavity, and larynx[1,2]. In addition, it includes information based on behavioral features of a speaker such as an accent, involuntary transforms of acoustic parameters, etc. As a result, speech production process differ among speakers. The rapid advancement of biometrics-based technology, measuring human behavioral or physiological characteristics for identifying and verifying an individual, piques technological interest on the mechanics of realizing human speech to ease/automate Human-Machine Interaction, which causes considerable research on the automatic speaker/ speech recognition process.

The application of face recognition that serves as a contactless method of identity verification was advised during the COVID pandemic. With the progression of technology, Voice recognition biometric systems hold great promise, heralding an era of advanced capabilities and transformative applications. This paper investigates the future trends that are expected to mould the landscape of biometric systems in the upcoming years. Innovative approaches that seek to improve accuracy, convenience, and privacy in the domain of digital security are investigated.

 The integration of biometrics with Artificial Intelligence (AI) and machine learning, holds the key to opening up new possibilities in personalized user experiences. Although the future of biometric systems seems promising and effective, it is important to consider the ethical and privacy challenges that result from their widespread implementation. As these systems handle and process sensitive biometric information, thus concerns regarding data protection, consent, and potential biases call for careful consideration. This purpose of this paper is to offer an insightful examination of the intriguing future trends in Voice recognition biometric systems, encompassing their potential applications in various industries, the impact of AI integration, and the ethical considerations surrounding their deployment. By exploring these emerging developments, we aim to throw light on the transformative potential of biometrics in remolding the digital security landscape and developing a secure, more user-centric world.

* **FUNDAMENTALS OF VOICE AS A BIOMETRIC**

\The speech signal conveys several levels of information. At the basic level, it delivers information through words. At other tiers, the message contains information such as language spoken, emotion, gender, and in general, the identity of a speaker. The principal goal of an automatic speaker recognition system is to extract, characterize and recognize the message/ information in the speech signal, and thereby impart the speaker's identity. The automated identification or extraction of a person’s personal identity by examining an individual spoken utterance is referred to as Speaker Recognition [3-6]. Speaker Recognition activity can be divided into two main tasks, Speaker Identification(SI) and Speaker Verification(SV). The primary distinction between identification and verification is the number of decisions is equal to the size of the population in the former, whereas, in the latter, there is acceptance or rejection, regardless of the population size. While Speaker recognition is regarded as a general organic process, the tasks of Speaker Identification and Verification are associated with the Speaker Recognition process. Speaker Identification is related to the determination a speaker’s identity whilst Speaker Verification refers to the task of validating a speaker’s claimed identity [7-9].

 In the case of forensic application of SV systems, the likelihood of obtaining sufficient data is lower in the enrolment process[10]. Access control-type cases limit the average utterance length to a few seconds at most[11]. Consequently, it is crucial to engage in research efforts to obtain reliable SV performance under short-duration conditions. The quest to develop an SV system suitable for real-world implementation has advanced significantly over time[12-14]. Thorough assessments on SV systems are presented in[1,2,15], considered overall issues and techniques in SV, and lists limited duration as one of the problems in ASV[16].

 With the development of numerous methodologies apt for modeling speaker characteristics the field of SV has experienced significant advancements in past decades. These developments in the domain of SV have created opportunities toward practical deployable systems for person authentication. Various initiatives were made to develop person authentication systems in realistic scenarios like Remote person authentication[17], smart home security applications [18], speech biometric attendance systems [19], verification over a telephone-based network[20] highlight SV with short utterances, as a developing field for having deployable systems with real-world applications.

 The short utterance-based scenario comes into the picture when person authentication is the basis for practical deployment as authentication a person requires relatively short amount of time with most of the other biometric attribute-based systems utilized in practice. The usage of short utterances for SV minimizes the required testing time while also offering comfort to the speakers by relieving them of the burden of speaking. In this regard, the text-dependent SV has been identified as a promising contender for having a deployable system as it requires less amount of train/test time. However, there is a great risk 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:

The initial phase of voice recognition involves the development of each user's voiceprint. A voiceprint is a distinctive representation of a person's vocal traits that contains information like pitch, tone, accent, cadence, and pronunciation. During enrollment, the person’s voice is captured, analyzed, and used to generate a template that will be stored in the system's database.

2. Feature Extraction:

The system utilizes the created voiceprint to determine its key characteristics. These characteristics are selected depending on its ability to reliably identify one person from another. Mel-frequency cepstral coefficients (MFCCs), LPCC, and x-vectors are commonly used in speech recognition systems as they effectively define the vocal attributes.

3. Pattern Matching:

During the verification or identification procedure the speech recognition system is engaged in recording the user's voice sample followed by extraction of its attributes. A comparison between the extracted features and the voiceprint saved in the database utilizing a pattern-matching process is done. The technology can confirm the speaker's identification only if the features matches the stored voiceprint.

4. Validation:

A threshold is established to validate the precision of the test and training data.

 The significant advancements in SV research are attributable to classifier domain improvements. The earliest SV systems originated using vector quantization (VQ)[21], a dynamic time warping[22-25] approach. Later, with the development of GMM[26], ASV research has progressed in the past two decades with greater emphasis on channel compensation[27], data variability, etc. GMM with the UBM[28] was presented as an enhancement over independently trained GMM using the maximum-likelihood(ML) approach[29]. The latent variable approach has ushered another new paradigm in ASV technology. For instance, factor analysis(FA)-based approaches were proposed to model the intersession variability in the context of GMM supervector[30]. Inspired by the success of joint FA(JFA), that is, speaker factors directly as features for classification, Dehak et al.[31] implemented single total-variability(TV) subspace-based modeling of the speakers, that differed from separate subspaces for speakers and channels in JFA. Recent SV technology have centred on TV modeling, also known as i-vector. Modeling the i-vector space utilizes a separate speaker and channel-dependent subspaces with Gaussian probabilistic LDA(GPLDA)[32], and the intersession variability is dealt with efficiently using this approach. With present-day ASV technology's i-vector method, variable-length speech utterances can be efficiently represented in fixed dimensions. Deep-learning-based approaches have recently garnered a great deal of attention and extensive interest in various domains[33]. The studies employed Deep Neural Network (DNN) models trained for voice recognition to generate UBM for SV, such as acoustic models, with rich data phones that may be applied to generate more effective background models [34–38]. DNNs are successfully adopted for speaker information features extraction.

 Voice recognition is becoming a popular and reliable biometric tool for speaker identification with distinctive speech characteristics as the basis. For purposes of identification and authentication, biometric systems are dependent on the distinctive individuals’ physical and behavioral characteristics. Voice recognition, another term for speaker recognition, is a type of behavioral biometrics that explores 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:

Voice recognition systems are classified into two categories: text-dependent and text-independent. Text-dependent systems require users are to utter a certain phrase or a group of phrases throughout the authentication procedure. These technologies are more safeguarded despite being less user-friendly. On the other hand, text-free systems are more user-friendly as they permit users to speak freely, but may be less secure.

* **FUTURE TRENDS IN VOICE RECOGNITION**

The research effort to address the issues of short utterances for SV systems has substantially intensifies in recent years[17, 39]. The problem is handled at many levels of the ASV framework's sub-systems, including feature extraction, speaker modeling, score normalization, and score calibration. The low performance of the conventional state-of-art speaker recognition models with short utterances resulted in a study of different pattern recognition approaches and normalization techniques,that are capable of dealing with limited data short utterances. These techniques can serve as a gateway to identifying improved recognition system. This research work is engaged in finding/exploring various procedures and algorithms to deal with limited data on short utterance-based SV.

 Integration of different biometrics improves the accuracy and reliability of the personal authentication system.

- Security and access control: Because each person's voice is unique and challenging to reproduce or fabricate, speaker recognition provides a strong and secure authentication technique. Speaker recognition offers real-time authentication that permits quick and seamless access control. Speaker recognition is a non-intrusive form of identification as it does not require physical contact, unlike some biometric technologies such as fingerprint scanning.

Speaker recognition can be utilized for remote verification over the phone or through voice-controlled applications, boosting the versatility of the authentication process.

- Forensic Speaker Recognition: Law enforcement utilizes speech biometrics to match speech samples obtained from crime scenes, such as threatening phone calls, ransom requests, or recorded messages left by suspects. Comparison of these speech samples to known voiceprints of prospective suspects or individuals of interest helps the investigators in extraction of information and limiting the list of suspects. Speaker recognition supports voice lineups akin to traditional police lineups in which eyewitnesses identify suspects from a lineup of images. Witnesses or victims who have heard the offender's voice can review the voice samples of various suspects to select the one they believe most closely resembles offender's voice. Voice biometrics can assist in determining whether 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: Amazon Alexa, Apple's Siri, Google Assistant, or Microsoft's Cortana, are popular Virtual voice assistants found in voice-controlled gadgets. These voice assistants employ speaker recognition to identify users and customize responses depending on their preferences and previous encounters. Individuals with hearing impairments may be able to utilize speaker recognition because it emphasizes speech qualities over listening abilities.

     Over the years, voice biometrics has had a broad range of applications eg. customize services or information by voice, Intelligence applications, finance, banking and surveillance, and criminal-forensic investigations. Despite the progress in speaker recognition technology, challenges persists, such as dealing with background noise, alterations in a user's voice due to factors like illness, and attempts to spoof or deceive the system. Ongoing research revolves around enhancement in accuracy, robustness, and adaptability to different speaking conditions.

* **CONCLUSION**

Voice recognition offers both benefits and drawbacks as a biometric technology. One of the obstacles includes variations in voice as a result of aging, health issues, or emotional events. Ambient noise and other factors may also affect the accuracy of voice recognition systems. However, there are several pros to voice recognition. It is a non-invasive biometric approach because it utilizes the user's speech only for authentication. The approach assists the users to easily authenticate their identities by speaking, making it user-friendly and convenient. In addition to other tasks, Voice biometrics may be used for forensic investigations, telephone-based authentication, and access control.

**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