Overview of Voice Biometric Systems: Voice Person Identification and Challenges

Ajimah N. E, Ezukwoke N., Dialoke I.C., Odaba A., Iloanusi O. N.
Department of Electronic Engineering, University of Nigeria, Nsukka
nnabueze.ajimah.pg80983@unn.edu.ng, nnabuike.ezukwoke@unn.edu.ng,
ikennadialoke@yahoo.com, alphaeus17@yahoo.com, ogechukwu.iloanusi@unn.edu.ng

Abstract- Recognizing persons through their voices has been one of the earliest things a human does, as a baby could identify the voice of their mother even as a foetus in the womb and it get intensively better after birth. Researchers have worked on seeing how machines can recognize individuals the same way man does. Audio Signal Processing (ASP), Machine Learning (ML), Matching Algorithm Techniques have been developed to bring about successful Automatic Voice Recognition (AVR). This work gives an overview of voice recognition in totality as it cut across voice recognition down to the physics/dynamics of voice production by a human. This paper discussed clearly the challenges voice recognition faces and the ways to mitigate this challenges.
Keywords: *Audio Signal Processing, Machine Learning, Automatic Voice Recognition.*

1.0 Introduction

The requirement of biometric authentication for humans in real world has brought about the ongoing research topic on the use of biometric technologies (Tsalakanidou, Malassiotis, & Strintzis, 2007). Voices are very important that we cannot have an effective communication without using our voices. They are means through which clear communications are carried out easily. Information about the state of a speaker's emotion can be pictured from a voice (Tovarek, Ilk, & Partila, 2018). Imagine colleagues in a workplace having a conversation with one another in a closed hall and another colleague of theirs who is outside the hall but listening to the conversation will not only predict the number of people that has spoken in that hall but can also give account of who owns a particular speech. A fingerprint as a biometric trait is just an image of ridges and furrows on the skin of fingers; a voiceprint, on the other hand, constitutes a combination of the person's accent, inflection and rhythm as well as physical factors that gives information of the size and shape of a person's nasal passage, length of the vocal tract, the diameter of the vocal cords, etc. where also the gender, the ethnicity, age and sometimes the body size of the speaker can be predicted from the information sieved from the voice signal (Latinus & Belin, 2011). The first major implementation of biometric identity verification was traced to Second World War, when the US soldiers used a machine called spectrographs to intercept voice transmission and to track their enemies. Spectrograph was primitive and erroneous as at then until the year 1976 when the Texas instruments invented the first modern voice biometrics machine with the capability of accurate registration and determination of an end user's voiceprint with a better precision (Maltoni, Maio, Jaint, & Prabhakar, 2009).

2.0 Reviewed Work on Voice Recognition System

Recognizing humans by their voices by the help of machines is almost impossible without the proper understanding of how our natural ears pick information. This impossibility has pulled the interest of millions of researchers across the globe to develop interest to knowing how these sounds are been produced and how there are perceived by the human ears. We have reviewed works by some researchers on their contributions towards the growth and development of voice recognition.

Singh and Verma, (M. Singh & Verma, 2011) worked on Speech Recognition using Neural Network, where they discussed the link between the biological neuron and how it is related to our artificial machine of today and neural network-based. Rani, et al. (Rani, Rani, & Kakkar, 2015), also worked on Speech Recognition using Neural networks, though their work was more elaborate than that of Manvendra Singh and Karmel Verma (M. Singh & Verma, 2011). Palaz et al., (Palaz, Magimai, & Collobert, 2015) Worked on Analysis of CNN-Based Speech Recognition System using Raw Speech as input. In the architecture, the neural network is fed with a sequence of raw input signal that was split into frames and outputs a score for each of the frame. Osamma et al., worked on Convolutional Neural Networks for Speech Recognition (Abdel-hamid et al., 2014). Though recently there was improvement in the performance of speech recognition using the Hidden-Markov-Model (HMM) over the conventional Gaussian-Mixture-Model (GMM) performance due to the ability of the Deep Neural Network (DNN) to model complex correlations in speech features Iosif et al., in their work Comparison of Speech Features on the Speech Recognition Task (Iosif, Todor, Mihalis, & Fakotakis, 2007) reviewed HMM-Based speech recognition system so also on speech parameterization techniques where he discoursed on some parameterization techniques used by some authors and also mentioned. Ahmad et al., (Ahmad, Helmy, Wahab, Verma, & Sinha, 2017) Proposed an Arduino-based intelligent mechanism that replaces a mechanical key in a motorcycle, by utterance of a word by the owner to 'ON' or 'OFF' it. Gofman et al. (Gofman, Smith, & Mitra, 2016), in their work on feature level fusion of biometrics on mobile devices. The author suggested that integration of features from multiple biometric modalities could improve recognition accuracy as they proposed a multimodal biometric using HMM that fused data from both face and voices. It was implemented in real-world operational condition using a Samsung Galaxy S5 (SG5). Singh et al. (R. Singh, Gencaga, & Raj, 2016) saw the need to voice mimicking under control then decided to understudy the voice impersonation of mimic experts, they paid attention to formant and formant-related extent to find out the extent and type of manipulations. Tandogan et al. (Tandogan, Sencar, & Tavli, 2017) measuring the uniqueness of human voices, the used the MFCC to extract the speaker's utterances which was modelled as a finite number of Gaussian distributions.

2.1 Voice Biometric and the Systems

Voice biometric is a technology that utilizes the nature in which sound is produced by humans, otherwise called voice biology, as a measure to differentiating people. The parameters that a quantified in the sound production systems are: how sounds are made, the length of the vocal tract, the width of the vocal cord. Measuring voice is as useful as measuring the fingerprint, since set of voice characteristics varies from person to person. There is no two individual that have exactly the same voiceprint. An algorithm has been developed that accepts voice samples and accepts characteristics from the voice by the Voice Biometric Group. The voiceprint is stored in the database, which is compared with the incoming voiceprint during authentication and identification. A Voice recognition system may be called voice authentication/verification or voice identification.

2.2.1 Voice Authentication/Verification

This is a voice technology that authenticates a person’s ID by comparing the captured voiceprint with her previously captured voiceprint reference template. This is a one-to-one matching of the query voiceprint with the voiceprint template in the database of the system to confirm how genuine the claimed identity is. Verification will either accept or reject people.

2.2.2 Voice Identification

This is a voice technology that recognizes an individual by searching the entire enrolled voiceprint in the database for a match. It conducts one-to-many comparisons to ascertain whether an individual is in the database or not and if so returns to the identifier reference that matched. The identification system does not need a claim from the subject.

2.2.3 Developing a Voice Identification System

In a Voice Identification System they are of two broad stages namely which are;

a. The Enrolment Stage and,

b. The Query Stage.

The ***Enrolment stage*** involves recording the live voice of the individual and saving it in the system's database whereas the ***Query stage*** involves the same process as the enrolment but this time the query template has to end its journey at the decision module where the system now has to compare the query voice with the stored template in the database. If there is a hit we have a match thence the person is Verified/Identified otherwise is the scenario if there is a miss. A concise knowledge of a Voice Recognition System could be drawn from Figure 1 as it carries detailed information about the recognition system. Voice being an analogue signal cannot be worked on by the machines so it has to pass through some stages which include Audio to Digital Converter (this happens at the recording stage i.e. from the microphone) that converts the analogue voice to digital voice. The digital audio could now be processed digitally and features from the voice are then extracted and stored in the database as a template during the enrolment stage or compared with the stored template during the query stage. Voice feature extraction is the bedrock of the voice recognition system as it decides how precise and robust the system could be which has led to so many researchers come up with the following voice feature extraction techniques; Mel-Frequency Cepstral Coefficient (MFCC), Linear Prediction Coefficient (LPC), Linear Prediction Cepstral Coefficient (LPCC), Line Spectral Frequency (LSF), Discrete Wavelet Transform (DWT) and Perceptual Linear Prediction (PLP).

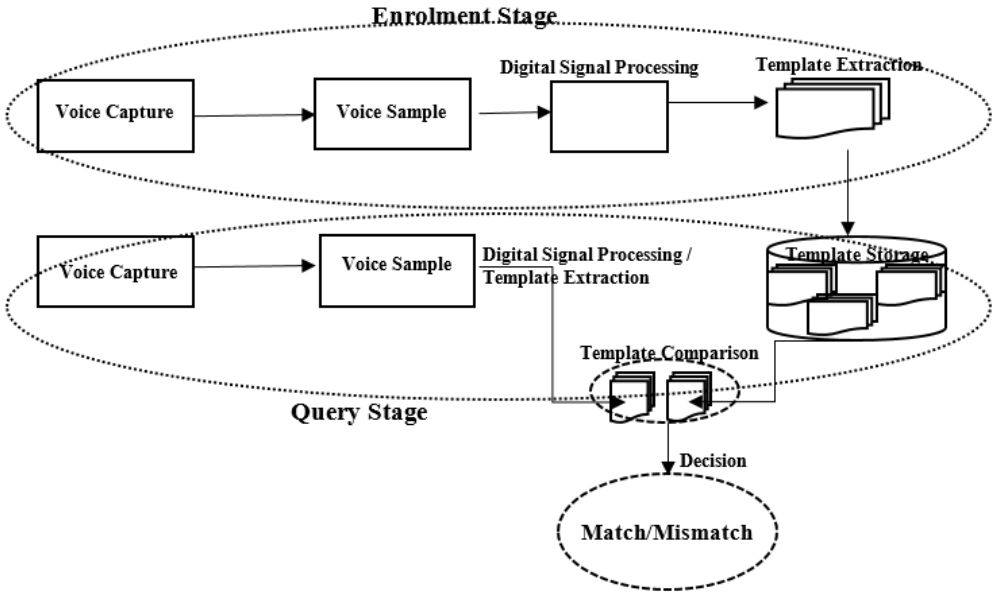

Figure 1: Block Diagram Showing the interaction of modules in voice Recognition System

3.1 Achieving an Efficient Voice Verification

Voice Verification System cannot be void of error due to contributory factors like background noise, device quality, age, ailment, and emotional state. In the next subsection, we will be discussing concisely these challenges as it affects the Voice Verification Systems (Abdulla & Yushi, 2010).

3.1.1 Challenges of Voice Verification System

Despite the advancement in technology of the voice biometric technology in automatically identifying humans by their voices, ambient temperature, stress, diseases medication, and other physical challenges can adversely impact this automated system(Latinus & Belin, 2011). Mentioned below are the list of the challenges that affect the performance of voice verification adversely:

i. **Background noise**: This is one of the biggest problems voice verification systems face as not only can it affect the matching performance but can also affect the standard of the template that was collected in such a noisy background(Paul & Saha, 2017).

ii. **Device Quality**: the quality of the devices used has a lot to contribute to the quality of the audio that is been recorded (Tovarek et al., 2018). Most high-end devices come with a qualitative microphone that has the power to produce very high quality sounds. The quality of a microphone is rated based on the following characteristics; sensitivity, frequency response, maximum sound pressure level.

iii. **Age of the Subject**: Human voice characteristics change as they advance in age. This implies that a database that is in use now will not be perfectly matching these people in years to come as so many of the children will then become adults and young men would be getting old with significant changes in their voices (Tovarek et al., 2018).

iv. **Ailment or Sickness**: Some common ailment like catarrh and cough causes human to lose their voices entirely and thus losing their biometric identity. People with such an ailment find it almost impossible to be accepted into a system where their voices were used as security (Tovarek et al., 2018).

v. **Emotional state**: Being in any emotional state such as, excitement and depression, has a way of giving entirely different voice characteristics hence a system would not be able to recognise or verify such a person for who he or she is, as it does not in any way corresponds with the voice of the person in the normal state (Tovarek et al., 2018).

vi. **Replay attack**: Replay attack on the voice system happens to be another serious challenge to an Automatic Voice Identification System (AVIS) as one can as easily collect voices of someone in the database and then play it to the microphone to gain access to the system (Tovarek et al., 2018).

3.1.2 Suitable Method of Voice data collection

In voice verification systems, the quality and manner of collection of this data it helps in improving the performance of the entire system. Voice data collection falls in the category of primary data acquisition since it is collected directly from the subject which could be collected through communication with the respondent in one form or another(Tovarek et al., 2018),(Poddar & Saha, 2017). Good voice data collection entails better performance of the system. The following are methods to achieving the best quality and better dataset;

i. **Multiple voice data per subject**: The voice passphrase should be repeated several times for text-dependent recognition system and a variety of groups of phrases are spoken by a single individual for the text-independent recognition system (Poddar & Saha, 2017).

ii. **Moderate voice tone**: The speaker should use his or her normal voice tone hence avoiding shouting or whispering, which has a way of affecting the speaker's voice characteristics (Poddar & Saha, 2017).

iii. **Noise-Free environment**: The voice sample is collected in a relatively quiet environment since noise interference will affect the quality of the voice dataset (Ahmad et al., 2017).

iv. **Device Quality**: The device (microphone) in use should be in good form, as this will minimize noises. Poor quality devices have the ability of introducing noises to the recorded dataset, hence affecting the performance matching of the system in totality (Poddar & Saha, 2017).

4.0 Discussion

Voice Identification System could be used to identify a person and can also be used to classify individuals based on their age, gender, health condition, emotion, and ethnicity. This categorization or classification cannot be achieved without some techniques like that of feature attraction coming into play. Voice Recognition has to do with Authentication and Identification. Authentication/Verification has to do with when a claim to something or to be someone the recognition that comes into play but in a situation where no claim is laid it becomes identification.

5.0 Conclusion

Voice recognition system is an important field of person identification as it is been deployed by manufacturers of several devices and machines such as; cars, mobile phones, Personal Data Assistant (PDA), bots. Voice has an advantage over several biometric modalities for its overtness and low cost of implementation but its greatest nightmare is noise which could come from the microphone or the immediate surrounding where the voice is been recorded.

References

Abdel-hamid, O., Mohamed, A., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional Neural Networks for Speech Recognition. *IEEE*, *22*(10), 1533–1545.

Abdulla, W. H. ., & Yushi, Z. (2010). Voice biometric feature using Gammatone filterbank and ICA. *International Journal of Biometric*, *2*(4), 330–349.

Ahmad, N., Helmy, M., Wahab, A., Verma, G., & Sinha, A. (2017). Motorcycle Start-stop System based on Intelligent Biometric Voice Recognition. In *Material Science and Engineering* (p. 6). http://doi.org/10.1088/1742-6596/755/1/011001

Gofman, M. I., Smith, N., & Mitra, S. (2016). Hidden Markov Models for Feature-level Fusion of Biometrics on Mobile Devices. *IEEE*, *5*, 0–1.

Iosif, M., Todor, G., Mihalis, S., & Fakotakis, N. (2007). Comparison of Speech Features on the Speech Recognition Task. *Journal of Computer Science*, *3*(8), 608–616.

Latinus, M., & Belin, P. (2011). Human voice perception. *ScienceDirect*, *21*, 143–145.

Maltoni, D., Maio, D., Jaint, A. K., & Prabhakar, S. (2009). *Handbook of Fingerprint Recognition* (Second). London: Springer-Verlag London.

Palaz, D., Magimai, M., & Collobert, R. (2015). Analysis of CNN-based Speech Recognition System using Raw Speech as Input. In *Conference proceedings of interspeech* (pp. 11–15).

Paul, D., & Saha, G. (2017). Synthetic speech detection using fundamental frequency variation and spectral features I. *ScienceDirect: Computer Speech and Language*. http://doi.org/10.1016/j.csl.2017.10.001

Poddar, A., & Saha, G. (2017). Speaker verification with short utterances : a review of challenges , trends and opportunities. *The Institution of Engineering and Technology*, *7*(2), 91 – 101. http://doi.org/10.1049/iet-bmt.2017.0065

Rani, P., Rani, S., & Kakkar, S. (2015). Speech Recognition using Neural Network, (Icaet), 11–14.

Singh, M., & Verma, K. (2011). Speech Recognition Using Neural Networks, *2*(March), 108–110.

Singh, R., Gencaga, D., & Raj, B. (2016). FORMANT MANIPULATIONS IN VOICE DISGUISE BY MIMICRY Language Technologies Institute , Carnegie Mellon University , Pittsburgh , USA. *IEEE*, *ii*.

Tandogan, S. E., Sencar, H. T., & Tavli, B. (2017). Towards Measuring Uniqueness of Human Voice. *IEEE*, 6.

Tovarek, J., Ilk, G. H., & Partila, P. (2018). Human Abnormal Behavior Impact on Speaker Verification Systems. *IEEE Access*, *6*, 40120–40127. http://doi.org/10.1109/ACCESS.2018.2854960

Tsalakanidou, F., Malassiotis, S., & Strintzis, M. G. (2007). A 3D face and hand biometric system for robust user-friendly authentication. *ScienceDirect*, *28*, 2238–2249. http://doi.org/10.1016/j.patrec.2007.07.005