Speaker Recognition System: Vulnerable and Challenges

Naufal Alee[#1], Phaklen Ehkan[#2], R.Badlishah Ahmad[#], Naseer Sabri[#]
[#]School of Computer and Communication Engineering
Universiti Malaysia Perlis (UniMAP)
Perlis, Malaysia
[1] thepsin@gmail.com
[2] phaklen@unimap.edu.my

***Abstract*— Recently speaker recognition system became high interesting by researchers for both software and hardware solutions. Different technologies have been adopted to implement speaker recognition system that has performance with optimal time response with acceptable accuracy. Research progresses are going on to provide highly durable and precise recognition system that can be embedded into critical implementation such as authentication for civilian or military aspect and also for future unmanned automation system. This paper is an introductory to speaker recognition system and its technology. Vulnerable of this system, its state in detail and propose solution has been shown.**

Keyword- Speaker Recognition; Biometric System; Gaussian Mixture Model (GMM); Field Programmable Gate Array (FPGA)

I. INTRODUCTION

Speaker recognition is the process of recognizing who is speaking by using characteristics of speaker's voice as a biometric base. It's also known as voice recognition. The known publications on speaker recognition have first published in 1954 [1]. Speaker recognition uses the acoustic features of speech that have been concluded to differ between individuals. These acoustic features reflect both anatomy (size and shape of the throat and mouth) and learned behavioral patterns such as voice pitch and speaking style. This incorporation of learned patterns into the voice templates has earned speaker recognition its classification as a "behavioral biometric". Speaker recognition can be categorized as speaker identification or speaker verification depends on the purpose of recognition either it's to identify speaker from a group of speakers or to verify the identity that claimed by the speaker [2].

The evolution of voice technology has significantly advanced using of artificial intelligence and technology of forensic science because it endows machines with the human-like abilities to distinguish people's identity from one another [3]. Speaker recognition technologies together with an accepted biometric feature nowadays offers high level of security, and these technologies currently are applying in many daily applications ranging from civilian to military and high central intelligent agency works. These include the access control system, security control for confidential information, transaction authentication as well as the telephone banking.

The success of speaker recognition system depends highly on how to classify a set of feature used to characterize speaker specific information [4] [5]. However, pattern classification from speech signal remains as a challenging problem encountered in general speaker recognition system, including speaker verification and speaker identification. Recent development in classifying speaker data from a group of speakers is still insufficient to provide a satisfying result in achieving high performance pattern classification. There are two main difficulties in pattern classification field; first, how to maintain accuracy under incremental amounts of training data and second, how to reduce the processing time as real time systems regarding efficiency and simplicity of calculation [6] [7].

II. SPEAK RECOGNITION SYSTEM: REVIEW

Biometric systems are the automated method of verifying or recognizing the identity of a person on the basis of physiological characteristic, such as a finger print, face pattern and human voice [8]. The human voice conveys information about identity of the speaker. Speaker recognition is the computing task of validating a user's claimed identity using characteristics extracted from their voices [9] [10]. Voice of a person has many prominent characteristics like pitch, tone which can be used to distinguish a person from the other. There are two main phases in speaker recognition system called *feature extraction phase* and *pattern classification*. The speaker recognition system has to process the speech signal in order to extract speaker discriminatory information from it. The purpose of *feature extraction* is to convert the speech waveform to some type of parametric representation at a considerably lower information rate. A wide range of possibilities exist for parametrically representing the speech signal for the speaker recognition task, such as Linear Predictive Coding (LPC) [11], Mel Frequency Cepstral Coefficients (MFCC) [12], Perceptual Linear Predictive (PLP) [13] and others. The MFCC and PLP are among the most popular acoustic features used in speaker recognition.

The *pattern classification* plays as an essential part in speaker modeling component chain. The results of it strongly affect the speaker recognition engine to decide whether to accept or reject a speaker. Early pattern classification was produced through Dynamic Time Warping (DTW) technique [14] [15] and Hidden Markov Models (HMM) technique [16]. These techniques are not really efficient for real time application due to characteristic of text dependent recognition. Vector Quantization (VQ) [17], Gaussian Mixture Model (GMM) and Support Vector Machine (SVM) [18] as the alternative methods were introduced for speaker recognition to solve the problem. The GMM is a density estimator and is one of the most commonly used types of pattern classifier. It has been demonstrated its effectiveness performances in text independent speaker identification [19]. The GMM technique of pattern classification in previous studies appeared to have several advantages. However, the process practically does not always produce satisfied result due to the long computational time [20] [21]. Consequently, alternative methods must be sought in order to reduce processing time problem for GMM technique. GMM performed very well performance, but its training process requires a lot of time and they get numerically unstable when trained with small amount of data. The main problem is the inversion of the covariance matrices. The use of GMM are most common due to the ability that it can be performed in a completely text independent situation. Besides, GMM are based on probabilistic framework and it provides high-accuracy recognition.

There are some hybrid methods for speaker pattern classification. They draw the attention of the researchers because it was proved with significant improvement for speaker recognition accuracy rates such as hybrid GMM with artificial neural network [22], hybrid GMM/VQ [23] and hybrid GMM/SVM [24, 25]. Most of these hybrid systems use GMM because it was able be performed in a completely text independent situation [26]. Performance of speaker recognition systems in term of accuracy rate has been significantly improved over hybrid conditions. However when speaker recognition is adopted in real-world application, processing time issue is often observed [27]. Meanwhile, current works for the hybrid production of speaker recognition are directed more towards accuracy problems, not processing time problems. Therefore, it is encouraging if a speaker recognition task can be conducted in a "good and fast" pattern classification machine such as in Field Programmable Gate Array (FPGA) based hardware implementation.

A. Speaker Recognition based on Hardware Technologies

FPGA platforms have been used to solve the speech problem focusing on speech recognition. Lin and Rutenbar [28] have been motivated to achieve a large speedup over time in order to accelerate searches of multimedia databases and demonstrated speedup of a 17 times over real time whilst maintaining good recognition accuracy. Yoshizawa et al. [29] aimed to achieve real-time recognition performance comparable to that of a standard microprocessor, but at much lower power dissipation and demonstrated a 10 times improvement in total energy dissipation over a system based on a TMS320VC5416 DSP for real time recognition tasks. Ramos-Lara et al. [30] have investigated the problem of hardware implementation of speaker identification, and do not aimed to achieve large speedups of performance, but instead to achieve identification using hardware at lower cost than a standard computer system. Compare with Pentium IV computer for a single voice stream, Xilinx Spartan 3 2000 FPGA achieved the same performance but using only 24% of the cost.

The hardware implementations initially tended to be based on parallel arrays of one kind or another, often using customize chips. As the technology improved, the focus has shifted towards serial implementations, making use once again of customize chips such as application specific integrated circuits (ASICs), microcontrollers or Digital Signal Processing (DSP) applications. Since the appearances of FPGA, that too has been applied as a platform for our work. ASICs customized for a particular use are very expensive even though they provide the highest performance. DSP-based designs, on the other hand, are cost efficient and low in power consumption and heat-emission. However, they only provide a limited speed for data processing because using special memory architectures that are able to fetch multiple data and/or instructions at the same time, they are susceptible to arithmetic saturation. FPGAs are usually slower than ASICs but have the advantage of shorter time to market, ability to be re-programmed in the field for errors correction and upgrades, flexibility, and reducing-cost. Therefore, they combine many advantages of ASICs and DSPs [31]. The use of hardware description languages (HDLs) allows FPGAs to be more suitable for different types of designs where errors and components failures can be limited. Due to the exponential increase of technologies, designers are faced with problems that require the advent of systems that can be fast, flexible, and mainly re-programmable. FPGAs because of their advantage of real-time in-circuit reconfigurability make the FPGA based system flexible, programmable, and reliable. They also facilitate the prototyping of complex electronic logic designs.

New generations of FPGAs now offer very high logic capacity and contain embedded Arithmetic Logic Units (ALUs) to optimize signal processing performance. The new generations of design tools enable developers to implement and develop complex systems within a reasonable time with the help of including libraries of common DSP functions. FPGAs have been used in many areas to accelerate algorithms that can make use of massive parallelism and improving flexibility. FPGAs are able to exploit pipelining and parallelism in a much

more thorough way that can be done with parallel computers using general-purpose microprocessors or a single standard processor [32].

B. Speaker Recognition based on Combination of Hardware and Software

Once the designing process is completed, it is necessary to undertake a thorough testing of the FPGA design. Testing will normally be undertaken at each stage of the FPGA development process which includes timing analysis, functional simulation and other verification methodologies. Once the design and validation process is completed, a binary file generated by the proprietary software is used to configure the FPGA device.

Although FPGAs offer many advantages, there naturally have some disadvantages. FPGA are slower than equivalent ASICs or other equivalent ICs which are cheaper. However, ASICs are very expensive and costly to develop. This means that the choice of whether to use an FPGA-based design should be made early in the design cycle and will depends on such items as whether the chip will need to be reprogrammed, or equivalent functionality can be obtained elsewhere, and at allowable cost. Sometimes manufacturers may opt for an FPGA design for early product (prototype) when bugs may still found, and then use ASICs when design is fully stable and reliable.

FPGAs are used in many applications. In view of the cost they are not used in cheap high volume products, but instead FPGAs find application in a variety of areas where complex logic circuitry needed, and changes may be anticipated. The applications cover a wide range of areas from equipment for video and imaging, to circuitry for aerospace and military applications, as well as electronics for specialized processing.

III. SPEAKER RECOGNITION SYSTEM VULNERABILITY

It has always been difficult to develop a robust speaker recognition system because speech signal is dynamic, varies by many factors, the complexity of biometric algorithms and the need of working in real-time. There have seen significant progress being made to deal with this problem using different techniques in the past two decades [33]. The problem of speaker recognition belongs to a much broader topic in scientific and engineering so called pattern classification. The goal of pattern classification is to classify objects of interest into a number of categories or classes [34]. The categories or classes here are referred to the individual speakers. One proposed solution is to ascertain and enhance GMM pattern classification approach via reconfigurable hardware implementation for speaker recognition. This pattern classification approach should be able to handle large speaker database in short time limit, whereas the accuracy rate is still maintained or even higher than the conventional GMM pattern classification technique. Shen and Reynolds [35] analyzed the NIST speech corpus evaluation set by using GMM and concluded that more than 5 minutes are used for identifying 100 sets of speaker data. Similar reviews have been done by [36]. The GMM computational time will dramatically increase when dealing with large set of data. Therefore, banking authentication systems often verify user identity instead of identify user voice with full set of data [35]. However, there are still needs on GMM speaker recognition system such as the access control system.

A survey is carried out to investigate suitable solution for reducing GMM technique processing time. [37] claimed by reducing learning data can improve training speed for speaker identification. Similarly, [38] revealed a speaker pruning algorithm for real time speaker identification which is based on reducing training data. Meanwhile, [39] declared decision tree approach is effective for solving large data problem which can divide the whole set of data into separable classes. Several hybrid pattern classifications had been conducted in order to obtain better accuracy rates for speaker identification system according to [24][25].

Currently, there are many existing speaker recognition system available but most of them are based on software. The problem with the software is that its implementation slow in time and is not suitable in practically used. Most implementations have been based on software with high performance of microprocessors [40]. This kind of solution is unacceptable in terms of cost, size and power consumption. Conventional GMM approach in recognizing people identity from speech signal is still insufficient to produce data. It is time consuming and requires heavy computations. The emergence of speaker recognition technologies require pattern classification engine for speaker recognition manage to process huge speaker data sets in limited time. Hence, the embedded FPGA-based GMM algorithm with less computational time and capable to work on huge dataset should be developed.

IV. DISCUSSION AND CONCLUSIONS

Due to above factors, a new proposed solution of GMM based on FPGA-based hardware method which takes the advantage of parallel processing classification approaches for text independence speaker identification is required. The reason is that GMM is effective and provide a stable accuracy while handling large speaker data. This proposed solution focuses on construction of embedded FPGA-based system leading to decrease processing time and result in higher accuracy for text independent speaker recognition. Base on the previous work survey in this paper, it has clearly shown that successful implementation of speaker recognition system need a detailed study and a serious estimation of problem constraints and variety, especially that the platform

solution must combine software and hardware issues, thus an optimization techniques must adopted for both to yield an efficient system.

REFERENCES

[1] I. Pollack, J.M. Pickett & W. Sumby, "On the Identification of Speakers by Voice," Journal of the Acoustical Society of America, 26, pp.403–406, 1954

[2] T. S. Rao and E. S. Reddy, "Multimodal Biometric Authentication Based on Score Normalization Technique," in Intelligent Informatics, A. Abraham and S. M. Thampi, Eds. Springer Berlin Heidelberg, pp. 425–434, 2013.

[3] J.A. Markowitz, "Voice Biometrics," Communications of the ACM, 43(9), pp. 66–73, 2000.

[4] J. Deng and Q. Hu, "Open Set Text-independent Speaker Recognition based on Set-score Pattern Classification," in IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings (ICASSP), vol. 2, pp. II–73–6, 2003

[5] J. Sorensen and M. Savic, "Hierarchical Pattern Classification for High Performance Text-independent Speaker Verification Systems," in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, pp. I/157–I/160, 1994.

[6] X. He and Y. Zhao, "Fast Model Selection Based Speaker Adaptation for Nonnative Speech," IEEE Transactions on Speech and Audio Processing, vol. 11, no. 4, pp. 298–307, 2003.

[7] W. M. Campbell, "Generalized Linear Discriminant Sequence Kernels for Speaker Recognition," in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, pp. I–161–I–164, 2002.

[8] S. Y. Kung, M. W. Mak, and S. H. Lin, "Biometric Authentication: A Machine Learning Approach" Prentice Hall, 2004.

[9] J.P. Campbell, "Speaker Recognition: A Tutorial," Proceedings of the IEEE, vol. 85, no. 9, pp. 1437–1462, 1997.

[10] S. Furui, "Fifty Years of Progress in Speech and Speaker Recognition," The Journal of the Acoustical Society of America, vol. 116, no. 4, pp. 2497–2498, 2004.

[11] B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," The Journal of the Acoustical Society of America, vol. 50, no. 2B, pp. 637–655, 1971.

[12] 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.

[13] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech," The Journal of the Acoustical Society of America, vol. 87, no. 4, pp. 1738–1752, 1990.

[14] H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 26, no. 1, pp. 43–49, 1978.

[15] J. Liu, Q. Cheng, Z. Zheng, and M. Qian, "A DTW-based Probability Model for Speaker Feature Analysis and Data Mining," Pattern Recognition Letters, vol. 23, no. 11, pp. 1271–1276, Sep. 2002.

[16] L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proceedings of the IEEE, vol. 77, no. 2, pp. 257–286, 1989.

[17] V. Radová and Z. Švenda, "Speaker Identification Based on Vector Quantization," in Text, Speech and Dialogue, V. Matousek, P. Mautner, J. Ocelíková, and P. Sojka, Eds. Springer Berlin Heidelberg, 1999, pp. 341–344.

[18] R. Solera-Ureña, D. Martín-Iglesias, A. Gallardo-Antolín, C. Peláez-Moreno, and F. Díaz-de-María, "Robust ASR using Support Vector Machines," Speech Communication, vol. 49, no. 4, pp. 253–267, Apr. 2007.

[19] D. A. Reynolds and R. C. Rose, "Robust Text-independent Speaker Identification using Gaussian Mixture Speaker Models," IEEE Transactions on Speech and Audio Processing, vol. 3, no. 1, pp. 72–83, 1995.

[20] Q. Y. Hong, S. Kwong, and H. L. Wang, "Optimization of Gaussian Mixture Model Parameters for Speaker Identification," in Genetic and Evolutionary Computation – GECCO 2004, K. Deb, Ed. Springer Berlin Heidelberg, 2004, pp. 1310–1311.

[21] D. A. R. Dr and W. M. C. Dr, "Text-Independent Speaker Recognition," in Springer Handbook of Speech Processing, P. J. B. Dr, P. M. M. Sondhi, and P. Y. (Arden) H. Dr, Eds. Springer Berlin Heidelberg, 2008, pp. 763–782.

[22] B. Xiang and T. Berger, "Efficient Text-independent Speaker Verification with Structural Gaussian Mixture Models and Neural Network," IEEE Transactions on Speech and Audio Processing, vol. 11, no. 5, pp. 447–456, 2003.

[23] J. Pelecanos, S. Myers, S. Sridharan, and V. Chandran, "Vector Quantization Based Gaussian Modeling for Speaker Verification," in 15th International Conference on Pattern Recognition Proceedings, 2000, vol. 3, pp. 294–297.

[24] S. Fine, J. Navratil, and R. A. Gopinath, "A Hybrid GMM/SVM Approach to Speaker Identification," in 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP), 2001, vol. 1, pp. 417–420.

[25] M. Liu, Y. Xie, Z. Yao, and B. Dai, "A New Hybrid GMM/SVM for Speaker Verification," in 18th International Conference on Pattern Recognition (ICPR), 2006, vol. 4, pp. 314–317.

[26] F. Hou and B. Wang, "Text-independent Speaker Recognition using Probabilistic SVM with GMM Adjustment," in International Conference on Natural Language Processing and Knowledge Engineering Proceedings, 2003, pp. 305–308.

[27] Y. S. Moon, C. C. Leung, and K. H. Pun, "Fixed-point GMM-based Speaker Verification over Mobile Embedded System," in Proceedings of the ACM SIGMM workshop on Biometrics methods and applications, New York, NY, USA, 2003, pp. 53–57.

[28] E. C. Lin and R. A. Rutenbar, "A Multi-FPGA 10x-Real-Time High-Speed Search Engine for a 5000-word Vocabulary Speech Recognizer," in Proceedings of the ACM/SIGDA international symposium on Field programmable gate arrays, New York, NY, USA, 2009, pp. 83–92.

[29] S. Yoshizawa, N. Wada, N. Hayasaka, and Y. Miyanaga, "Scalable Architecture for Word HMM-based Speech Recognition and VLSI Implementation in Complete System," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 53, no. 1, pp. 70–77, 2006.

[30] R. Ramos-Lara, M. Lopez-Garcia, E. Canto-Navarro, and L. Puente-Rodriguez, "SVM Speaker Verification System Based on a Low-Cost FPGA," in International Conference on Field Programmable Logic and Applications (FPL), 2009, pp. 582–586.

[31] E. Ayeh, K. Agbedanu, Y. Morita, O. Adamo, and P. Guturu, "FPGA Implementation of an 8-bit Simple Processor," in IEEE Region 5 Conference, 2008, pp. 1–5.

[32] Sumedh, S.J. and Bhoyar, C.N. (2012). "FPGA Based Embedded Multiprocessor Architecture," International Journal of Scientific and Engineering Research, Volume 3, Issue 1, 1-6.

[33] S. Furui, "Recent Advances in Speaker Recognition," Pattern Recognition Letters, vol. 18, no. 9, pp. 859–872, Sep. 1997.

[34] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification. John Wiley & Sons, 2012.

[35] W. Shen and D. Reynolds, "Improved GMM-based Language Recognition using Constrained MLLR Transforms," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2008, pp. 4149–4152.

[36] P. Bruneau, M. Gelgon, and F. Picarougne, "Parsimonious Reduction of Gaussian Mixture Models with a Variational-Bayes Approach," Pattern Recognition, vol. 43, no. 3, pp. 850–858, Mar. 2010.

[37] T.-S. Lee, H.-J. Choi, S.-H. Choi, and B.-W. Hwang, “A Method on Improving of Enrolling Speed for the MLP-Based Speaker Verification System through Reducing Learning Data,” in PRICAI 2002: Trends in Artificial Intelligence, M. Ishizuka and A. Sattar, Eds. Springer Berlin Heidelberg, 2002, pp. 619–619.
[38] T. Kinnunen, E. Karpov, and P. Fränti, “A Speaker Pruning Algorithm for Real-Time Speaker Identification,” in Audio- and Video-Based Biometric Person Authentication, J. Kittler and M. S. Nixon, Eds. Springer Berlin Heidelberg, 2003, pp. 639–646.
[39] X. Wang, Z. Shi, C. Wu, and W. Wang, “An Improved Algorithm for Decision-Tree-Based SVM,” in The Sixth World Congress on Intelligent Control and Automation (WCICA), 2006, vol. 1, pp. 4234–4238.
[40] P. EhKan, T. Allen, and S. F. Quigley, “FPGA Implementation for GMM-based Speaker Identification,” International Journal of Reconfigurable Computing., vol. 2011, pp. 3:1–3:8, Jan. 2011.