Real-Time Implementation of Isolated-Word Speech Recognition System on Raspberry Pi 3 Using WAT-MFCC

Mohamed Walid, Souha Bousselmi, Karim Dabbabi and Adnen Cherif

Faculty of Sciences of Tunis, Tunis-El Manar University, 2092 El-Manar-Tunis, Tunisia

Summary

The automatic speech recognition (ASR) is a field of research that has been emerged in the early 1950s, and has been used in the literature for convenient and efficient human-machine interaction. For voice command system, it is based on implementation of isolated-word speech recognition and it can include many applications, such as voice-activated devices, robots, access control system, etc. In this paper, such isolated-word speech recognition system has been implemented on Raspberry-Pi 3 (RPi 3) board by combining Wave Atoms Transform (WAT) approach and Frequency-Mel Cepstral Coefficients (MFCC) (WAT-MFCC) with Support Vector Machine (SVM). The experiments have been tested on Arabic words database and the achieved results have proven the reliability of the proposed WAT-MFCC-SVM recognition approach with a rate of 100% and a Real-Time Factor (RTF) of 1.50.

Key words:

Isolated-word speech recognition, Real-Time implementation, WAT, MFCC, WAT-MFCC, SVM, Raspberry PI 3 board.

1. Introduction

Speech is an easier communication way for humans to express their ideas and feelings. In fact, using it as a means to control one’s environment is usually an intriguing thing. For this reason, the research in automatic speech recognition (ASR) undergoes increased and renewable increments. Indeed, several researches have been carried out during last decades in order to design such an ideal speech recognition system which is able to understand isolated-words speech in real-time, from different speakers and in various environments. Nevertheless, reaching this ultimate goal is still a persistent requirement for the recent developed ASR systems. Also, this task is more challenged due to the presence of large variations in speech signals such as, absence or lack of clear boundaries between words or phonemes, and the presence of undesirable noise signals caused by the variability of the speakers and their surroundings (e.g. gender, speed of speech, speaking style, and dialects [1, 2]).

There are many applications of ASR systems which have been released to accomplish different tasks ranging from the simplest to the most complex, such as air traffic control, speech-to-text input, ticket reservations, gaming, security and biometric identification, automobile sectors, home automation [3, 4]. Furthermore, the advancements recording in ASR research area have their good impact in life of disabled and elderly persons by offering them highly quality of assistance.

In the literature, there are various perspectives from which the ASR tasks were examined. In [5], some challenges of ASR were discussed and a brief over view on a number of well-known approaches was also presented. Indeed, two feature extraction techniques were considered by the authors in that work: the Mel-frequency cepstral coefficient (MFCC) and predictive linear coding coefficient (LPC), besides five other classification methods: knowledge-based approaches, templates-based approaches, artificial neural networks (ANNs), hidden Markov models (HMMs) and dynamic time warping (DTW). Thus, a comparison between many ASR systems was performed on the basis of the extracted features and classification techniques. Moreover, numerous approaches have been cited in [6] and employed as techniques in both pre-processing and feature extraction stages of an ASR system. In [7], the authors have presented different viewpoints for constructing ASR systems. In fact, they have considered that these systems are composed of a number of processing layers since several components are required, resulting in a number of their computations. Also, authors have concluded that the reduction in the present error rates of ASR can be obtained when choosing wisely the corresponding processing layers. In [8], both ASR and text-to-speech (TTS) research areas have been discussed by authors. In ASR section, they have considered different aspects for the classification of speech, such as cepstrum-based feature extraction techniques, data compression, and HMMs. Also, they have discussed different ways to increase robustness against noise. In [2], a discussion in the field of ASR from the perspective of pattern recognition was presented.

Mainly, ASR system consists of four phases: pre-processing phase, feature extraction phase, classification phase, and a language model [9]. In the pre-processing phase, the speech signal is transformed in order to further extract from it the consistent information in the feature extraction phase. In fact, there are common

Manuscript received March 5, 2019
Manuscript revised March 20, 2019

functions between pre-processing and feature extraction phases, such as, the pre-emphasis, framing, normalization, noise removal, endpoint detection [8, 10, 11]. A number of predefined features are then extracted from the processed speech in feature extraction phase. Thus, these extracted features have to be able to discriminate between classes while preserving the robustness to any external conditions, such as noise. In [12-14], it has been proved that the performance of ASR systems relies sharply on the selected feature extraction method since the used classifiers in the classification stage have to classify efficiently these extracted features. In [16, 3], various feature extraction methods have been proposed, such as, MFFC, LPC, and Discrete Wavelet Transform (DWT). For the language model, it is composed of different kinds of knowledge related to a language, such as the semantics and the syntax [15].

In recent decades, the creation of an ideal ASR system that has the ability to understand continuous and discrete speech in real-time (in any environment and from different speakers) was the center of interest of many research which still far from achieving this ultimate objective [16, 17]. Among the real-time applications of ASR, we can mention the security systems, automation and robotics, and so on. Most speech-related applications are classification-based applications. HMM and ANN models are frequently used for classification in ASR systems. The HMM model has the drawbacks that the probability function is the only one function that it can be used and the neighbor frames should be independent [18]. However, these constraints are resolved in ANN model as each neuron in the hidden layers has its activation function. For pattern recognition, ANN algorithm was considered as a good and a highly efficient classifier [18]. In subordinate database containing a vocabulary of 571 words, the Non Negative Matrix (NNM) had let to achieve a word exactness of 94% against 88% for HMM model using a low-perplexity language structure. In contrast, the word exactness has reached 58% for NNM against 49% for HMM without a language structure. Using TIMIT database, poor results have been reached employing both classifiers, but with a slight improvement for NNM model [19]. In Chinese speech recognition, the introduction of Deep Neural Network (DNN) has contributed to reduce the character error rate by 20% and it has outperformed Gaussian Mixture Model (GMM) in terms of performances [20]. In noisy conditions, more effective results have been obtained using DNN and support vector machine (SVM) classifiers in comparison with those of the state-of-the-art [21]. In [22], it has been shown that artificial neural network (ANN) is good for short or isolated word recognition. This is due to the fact that the reconfigurable hardware is faster than software and the neural network can offer a more improvement in term of speed when implementing it on hardware architecture. This is the case for image processing in robotics and pattern recognition applications when lesser time and fast time response, were achieved using neural network implemented on FPGA [23]. Only, 33% of FPGA resources have been used in this configuration. Although the neural network is slower on FPGA than on PC, but it is more stable, not dependant on the operating system and it is less expensive [24]. By using the neural network, the made vehicles become more intelligent. In [25], a hybrid approach destined for speech recognition system has been evaluated on Xilinx. This approach was designed on the basis of Multi-layer Perception (MLP) and has shown significantly reduction in power and area. In [26], the vector feature dimensions have been reduced using Self Organizing Feature Map (SOFM). Perceptual Linear Predictive (PLP), MFCC, and DWT represent the main features which have been extracted in this approach. Despite of the large reduction in the feature vector dimensions, the recognition accuracy was the same as that obtained with the conventional methods [26]. In [18], an isolated spoken word speech recognition system has been implemented on RPi 2. A neural network exploited as classifier and MFCC as feature are the main components constituting this system. In fact, an accuracy of 100% has been obtained on TIDIGITS corpus for speaker dependant speech recognition. However, it was lower in the case of speaker independent speech recognition. In [27], an implementation of an embedded isolated word recognition system (IWR) has been carried out on STM32F4-Discovery platform. The iteration of the system was repeated three times and was done in two different scenarios: an Anechoic Chamber that has a very high SNR and in a Normal Environment. An overall average rate error per word (WER) of 1.04% and 2.81% were found for the respective scenarios. The reported real time factor (RTF) was 1.43, which is in line with the level of performance reported in the literature.

Most of the solutions proposed in the literature that address the problem of Automatic speech Recognition (ASR) in embedded systems, are based on Digital Signal Processing (DSP) [28-32] or in FPGA [33-35]. We can also find some implementations based on Microcontrollers. Most of them require some communication channel with a remote server that processes the collected data and performs the real recognition [36-38]. Although these methods are powerful, they have higher latencies and consumptions.

In turn, there are some approximations of ASR systems which are totally based on microcontrollers. In general, these applications use simpler acoustic characteristics in order to reduce the computational cost and they range from very simple vectors [39-41] to some more complex and robust [42, 43].

Finally, other less popular implementations rely on dedicated chips to carry out the recognition process [44-46].

In this paper, we have implemented an isolated-voice recognition system on RPi 3 board. The proposed system was tested in different noisy conditions on Arabic database using Wave Atoms Transform (WAT) and MFCC as extracted features followed by SVM, MLP, and HMM as classifiers. WAT is one type of the wavelet-based transforms which also include wavelet packet transform (WPT) and DWT [47]. Furthermore, it is considered as a geometric tool used to analyze the signal in presence of noise and uncertainly that provide multi-resolution with multi-scale tools [48]. Moreover, Time-domain filtering is a simple denoising method which has been applied for corrupted signals [49] in order to remove high-frequency noise in low-frequency signals. However, this method cannot provide satisfactory results in real world conditions. To deal with this issue, we have proposed WAT-MFCC approach to more improve the real-time performances of isolated-word recognition system in different noisy conditions.

The remaining sections of this paper are given as follows: In section 2, the proposed method is presented. The real-time implementation of the proposed approach on RPi 3 board is exhibited in section 3. Discussion and analysis of different results are given in section 4. Conclusion and perspectives are drawn in section 5.

2. The Proposed Method

A comprehensive speech recognition system depending on WAT and SVM has been developed and implemented on Raspberry PI 3 board in order to increase the recognition accuracy in real-time. This system includes several stages that are illustrated in Fig. 1.

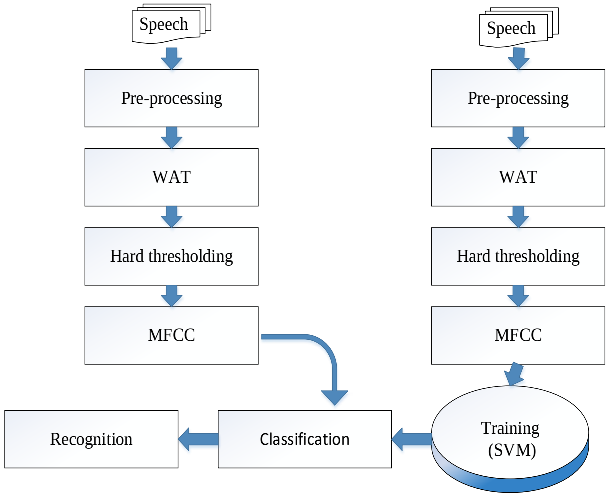

Fig. 1 Block diagram of the proposed speech recognition system.

2.1 Pre-processing Speech stage

The pre-processing consists of three phases: pre-emphasis, frame blocking, and windowing.

For the first phase, it is destinated to reduce noises of speech signal during the capture moments and smooth spectral form of its frequencies. Indeed, the expression of the pre-emphasis filter in time domain is given as follows:

$$?y\ x) = x(n) - ax(n-1) \qquad (1)$$

Where, a can be defined as a pre-emphasis filter constant which ranges between 0.9 and 1.0. Concerning the second phase, it is the frame blocking in which the audio signal is split into many overlapping frames in order to avoid finding a single deletion of signals. In fact, all signals during this process have to get into one or two frames. To perform this purpose, short-time analysis can be applied.

The third phase in the pre-processing stage is the windowing which can be described as an analysis process for long sound signals. Indeed, a Finite Impulse Response (FIR) digital filter is applied in this process in order to remove the aliasing signal forms caused by the discontinuities of the signal pieces which occur after the application of frame blocking process.

2.2 Feature Extraction Stage

2.2.1 Wave Atoms Transform (WAT)

The first step of our approach consists in decomposing the speech signal using Wave Atoms Transform (WAT). The particularity in this transform is represented by its capability to convert the temporal representation of a signal into a time-frequency one. Also, this domain transformation can reduce the redundancies and decorrelate the signal samples. Thus, better bit rates of transmission can be reached. Indeed, the WAT process is a technique which can concentrate speech information into a few coefficients [50]. Therefore, many coefficients will either be zero or have negligible magnitudes.

2.2.2 Thresholding

Thresholding is the main step in speech recognition in noisy environment; it allows the rejection of the coefficients in which the WAT transform is inferior to a given threshold. There are several methods of thresholding, such as the hard thresholding and the soft thresholding which are the commonly used methods. In this paper we have used the hard threshold given by the following equation:

$$C_{\mathrm{Re}} = \begin{cases} C_{\mathrm{Re}} \; if \; |C_{\mathrm{Re}}| \geq T \\ 0 \quad otherwise \end{cases} \tag{2}$$

2.2.3 Mel-Frequency Cepstral Coefficients (MFCCs)

The Mel-Frequency Cepstral Coefficients (MFCCs) are widely used in automatic speech recognition systems [51] due to their low complexity estimation and their good performance. It has been demonstrated that the MFCC representation approximates the structure of human auditory system better than the traditional linear and predictive features. However, MFCC coefficients are easily affected by common frequency localized random perturbations, to which human perception is largely insensitive. For each frame of speech signal, a MFCC vector is computed as follows: the power of the spectrum of a windowed signal block is mapped onto the Mel scale using triangular filters. The logarithm of the filter bank output is then again transformed by applying a Discrete Cosine Transform (DCT). The relationship between scale Mel and frequency is given by the following expression:

$$F\left(\mathrm{Mel}\right) = 2595 * \log 10\left(1 + f/700\right) \tag{3}$$

Indeed, an illustration about the computing of MFCCs coefficients is shown in Fig.2.

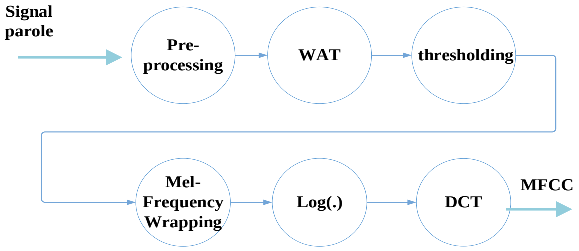

Fig. 2 Computation of MFCC Coefficients.

2.3 Classification

2.3.1 Support Vector Machine (SVM)

SVM classifier is a simple and an efficient computation of machine learning algorithms and it can be used to perform binary classification. Exploiting kernel function to transform the original input set into a high-dimensional feature space represents the main idea behind the use of this classifier. Therefore, that transformation can be used to solve non-linear problems [52]. For classification problems and patter recognition, SVM classifier has been widely applied and in limited training data, this classifier has outperformed other algorithms by reaching good classification performance [52].
The resulted hyper plan from the data separation of two groups is defined by the following expression:

$$f\left(x\right) = w^{t}x + b \tag{4}$$

Where b is a scalar, x is an input factor and w is an m dimensional vector.
SVM determines a hyperplane that corresponds to f(x) =0 for linearly separable data. However, the input samples are mapped into a high-dimensional feature space using φ function for non-linearly separable case, as follows:

$$f\left(x\right) = w^{t}\varphi\left(x\right) + b = 0 \tag{5}$$

and the decision function is consequently defined as:

$$f(x) = \mathrm{sign}\left(w^{t}\varphi\left(x\right) + b\right) \tag{6}$$

Recent studies aim to extend SVM to multi-class classification, which is often applied in real problems. Thus, the reduction of multiclass problem to a composition of several biclass hyper plane by drawing borders between classes [53]. In fact, there are two most popular methods for multiclass classification which are "one-against-one" (OAO) and "one-against-all" (OAA) techniques. For OAA approach, it is relies on “winner takes all” strategy, which aims to construct one SVM per class. That to say that to classify m classes, the one-against-all has to construct m binary SVM classifiers. This method is less sensitive to the imbalanced datasets, whereas it is more computationally expensive.
In this paper, we have applied "OAA" technique on Raspberry Pi 3 (RPi3) board using SVM library from python platform [54].

3. Real-Time Implementation on Raspberry Pi 3 (RPi 3) Board

3.1 Database

The 2/3 rd of the database has been consecrated for training, while the rest has been used for testing. We have tested the proposed model on Arabic database containing 11 Arabic spoken words of different directions (i.e takadam, tarajaa, 5alfa, amam, asraa, istader, sir, waraa, tawakaf, yamine, yassar). These words have been repeated 10 times at different levels of SNR (ranging from -5 to 20 dB).

3.2 Materials

3.2.1 Hardware

Raspberry Pi (RPi) can be defined as a single Linux board computer which has been extended to RPi 2 and RPi 3 versions. The general architecture of RPi 3 board model B is shown in Fig.3.

Fig. 3 The General Architecture of RPi 3 Board Model B.

In this work, we have used RPi 3 since its major advantages, such as its high speed which can be 50% faster [55] than RP2 in virtue of its processor (1.2 KHz), good memory capacity of RAM (1GB), and the extensible memory of external SD card. Compared to RPi 2, RPi 3 board includes wireless connections by integrating WIFI and Bluetooth, which makes it promised for the internet of Things (IoT) applications. An illustration about the main technical specifications of RPi 3 board is given in Table 1.

Table 1: Technical Specifications of RPi 3 Board Model B [55].

Feature	Type
CPU	1.2 GHz 64-bit quad core ARM Cortex-A53
Memory (SDRAM)	1 GB (shared with GPU)
USB 2.0 Ports	4 (5 with the on-board 5-port USB hub)
Video input	15-pin MIPI camera interface (CSI) connector , used with the Raspberry Pi camera or Raspberry Pi NoIR camera
Video outputs	HDMI (rev 1.3), composite video (3.5 mm TRRS jack)
On-board storage	Micro SDHC slot
On-board network	10/100 Mbit/s Ethernet, 802.11n Wireless, Bluetooth 4.1
Power source	5 V with Micro USB or GPIO header
CPU	1.2 GHz 64-bit quad core ARM Cortex-A53

3.2.2 Software

Raspberry PI 3 Model B can support many coding languages, such as C++, Java, Python, and…so on. In our work, we have used Python for the implementation of the proposed model since it makes the data processing faster than other programming language, as well as it contains rich toolbox.

4. Results and Analysis

In clean condition, the combination of SVM algorithm with WAT-MFCC feature has been contributed to get the best performances in terms of recognition accuracy (100%) and Real-Time Factor (RTF=1.50) compared to those obtained using HMM and MLP algorithms. Also, the use of WAT-MFCC as feature has led to improve the RTF compared to the use of MFCC only (1.50 and 1.70, respectively) with the preservation of high values of accuracy in all tests. Indeed, the obtained results for accuracy and RTF performances are summarized in Table 2.

In this paper, we have to mention that we have only included the results of RTF for the tested algorithms in clean condition as we have not got a significant improvement in noisy one.

Table 2: Recognition and RTF results for different speech recognition algorithms in clean condition.

Algorithm	Recognition Accuracy (%)	RTF
MFCC-SVM	100	1.70
MFCC-HMM	90.90	2.10
MFCC-MLP	93.93	2.70
WAT-MFCC-SVM	100	1.50
WAT-MFCC-HMM	93.93	1.87
WAT-MFCC-MLP	96.96	2.49

Table 3 illustrates the recognition results obtained with different speech recognition algorithms. From this table, we can remark that SVM has succeeded to get the best recognition accuracies for all levels of SNR and in all noisy conditions. This record has been achieved in comparison to HMM and MLP algorithms. Indeed, the best recognition accuracy (100%) has been obtained at 15 and 20 dB of SNR. For MLP and HMM algorithms, they have always led to achieve the best second and third recognition accuracies, respectively, at different levels of SNR in different noisy conditions. The best ones were 96.96% and 93.93% using MLP and HMM algorithms, respectively, in “cafe” noise at 20dB of SNR. Also, we can remark that the more the SNR level increases the more the recognition accuracy increases. In all tests, the worst results were recorded in “Traffic jam” noise in comparison with those obtained in “car” and “cafe” noises.

Table 3: Recognition results for different speech recognition algorithms in noisy conditions

Type of Noise	***Speech recognition algorithm***	-5db	0db	5db	10db	15db	20db
Traffic ***jam***	WAT-MFCC SVM	**63.63**	**72.72**	**87.87**	**90.90**	**96.96**	**100**
	WAT-MFCC HMM	54.54	60.6	63.63	69.69	75.75	75.75
	WAT-MFCC MLP	60.6	63.63	66.66	72.72	78.78	84.84
Car	WAT-MFCC SVM	**66.66**	**72.72**	**84.84**	**90.90**	**96.96**	**100**
	WAT-MFCC HMM	60.60	69.69	72,72	72.72	75.75	78.78
	WAT-MFCC MLP	63.63	70.70	72,72	78.78	84.84	87.87
Cafe	WAT-MFCC SVM	**72.72**	**72.72**	**75.75**	**84.84**	**100**	**100**
	WAT-MFCC-HMM	60.60	66.66	69.69	75.75	87.87	93.93
	WAT-MFCC-MLP	63.63	63.63	72.72	78.78	90.90	96.96

From Tables 2 and 3, we can say that the WAT-MFCC-SVM has proved its efficiency in terms of real-time performances in both clean and noisy conditions compared to HM and MLP algorithms. Furthermore, SVM algorithm has taken advantage from the combination of WAT approach with MFCCs coefficients which has sharply contributed to reach these performances by decreasing the amount of noise in test conditions.

Table 4 gives an illustration of some works which have been performed in real-time using isolated-voice speech recognition systems. We can summarize from this table that our proposed model has succeeded to obtain good real-time performances in comparison to other works which have been performed on different databases and in different noisy conditions.

Table 4: Real-Time performances obtained with other databases.

Database		Feature	Classifier	Accuracy (%)	RTF
Japaneese [56] (10dB)		MFCC	HMM	93.49 (Babble Noise)	2.37
		MFCC +NS		94.84 (Hfchannel Noise)	2.38
Google Speech [57]				Normal environment	
	Bangala	Google speech reorganization engine.		87	
	English			79	
				Calm environment	
	Bangala			85	
	English			90	
				Room environment	
	Bangala			86	
	English			76	
				Wind environment	
	Bangala			70	
	English			70	
Nepali database [58]		MFCC	HMM	75	-
Indian database [59]		MFCC	DTW	88	-
Lithuanian database [60]				97.70	-

5. Conclusion and Perspectives

In this paper, an implementation of isolated-voice speech recognition system has been proposed. In this system, the combination of WAT and MFCC with SVM as classifier has led to reach the best performances in both clean and noisy conditions compared to HMM and MLP algorithms. These performances are in terms of recognition accuracy and RTF in which this combination has reached 100% and 1.57s, respectively.

As further work, we suggest to test the proposed approach on-line or off-line on other databases, such as TIMIT database. Also, this approach should be implemented on other hardware architectures (i.e. FPGA, Ardouino…etc.) so that we can more view the progress of its performances in real-Time.

Acknowledgment

The authors would like to express their cordial thanks to Pr. Adnen Cherif for his valuable advice.

References

[1] Forsberg, M: "Why is speech recognition difficult?," Chalmers University of Technology, 2003, http://www.speech.kth.se/~rolf/gslt_papers/MarkusForsberg.pdf.

[2] O'Shaughnessy,D,"Invited paper: automatic speech recognition: history, methods and challenges," Pattern Recognition, 2008, 41, (10), pp. 2965–2979.

[3] Raman, S., "A discrete wavelet transform based approach to Hindi speech recognition," Int. Conf. on Signal Acquisition and Processing, 2010 (ICSAP'10), Bangalore, 2010, pp. 345–348.

[4] Junior, S.B., Guido, R.C., Chen, S., Vieira, L.S., Sanchez, F.L, "Improved dynamic time warping based on the discrete wavelet transform," Ninth IEEE Int.Symp. Multimedia Workshops, 2007 (ISMW'07), Taichung, Taiwan, pp. 256–263.

[5] Vimala, C., Radha, V,"A review on speech recognition challenges and approaches," World Comput. Sci. Inf. Technol., 2012, 2, (1), pp. 1–7.

[6] Anusuya, M., Katti, S,"Front end analysis of speech recognition: a review," Int. J. Speech Technol., 2011, 14, (2), pp. 99–145.

[7] Morgan, N,"Deep and wide: multiple layers in automatic speech recognition," IEEE Trans Audio Speech Lang. Process., 2012, 20, (1), pp. 7–13.

[8] O'Shaughnessy, D,"Interacting with computers by voice: automatic speech recognition and synthesis," Proc. IEEE, 2003, 91, (9), pp. 1272–1305.

[9] Cutajar, M., Micallef, J., Casha, O., Grech, I., & Gatt, E. (2013),"Comparative study of automatic speech recognition techniques," IET Signal Processing, 7(1), 25–46.

[10] Mporas, I., Ganchev, T., Siafarikas, M., Fakotakis, N," Comparison of speech features on the speech recognition task," J. Comput. Sci., 2007, 3, (8), pp. 608–616.

[11] Saha, G., Chakraborty, S., Senapati, S,"A new silence removal and endpoint detection algorithm for speech and speaker recognition applications," Proc. NCC 2005, 2005.

[12] Zamani, B., Akbari, A., Nasersharif, B., Jalalvand, A, "Optimised discriminative transformations for speech features based on minimum classification error," Pattern Recognit. Lett., 2011, 32, (7), pp. 948–955.

[13] Vimal Krishnan, V.R., Babu Anto, P,"Features of wavelet packet decomposition and discrete wavelet transform for malayalam speechrecognition," Recent Trends Eng., 2009, 1, (2), pp. 93–96.

[14] Alkhaldi,W., Fakhr,W., Hamdy, N,"Automatic speech recognition in noisy environments using wavelet transform,"2002. Available from: http://www.wseas.us/e-library/conferences/skiathos2002/papers/447- 231.pdf.

[15] Jurafsky,D, Martin, J.H," Speech and language processing," (Prentice-Hall, 2009).

[16] Vimal Krishnan, V.R., Babu Anto, P,"Feature parameter extraction from wavelet subband analysis for the recognition of isolated malayalam spoken words," Int. J. Comput. Netw. Secur., 2009, 1, (1), pp. 52–55.

[17] Hennebert, J., Hasler, M., Dedieu, H,"Neural networks in speech recognition," Sixth Microcomputer School, Prague, Czech Republic, 1994, pp. 23–40.

[18] Patange, P. P., & Alex, J. S. R. (2017),"Implementation of ANN based speech recognition system on an embedded board," 2017 International Conference on Nextgen Electronic Technologies: Silicon to Software (ICNETS2).

[19] Steve Renals,David McKelvie and Fergus McInnes, "Comparative Study of Continuous Pattern "Recognition Using Neural Networks and Hidden Marcov Model ," IEEE 1991.

[20] Xiangang Li, Yuning Yang, Zaihu Pang, Xihong Wu, "A comparative study on selecting acoustic modeling units in deep neural networks based large vocabulary Chinese speech recognition," Neurocomputing, Volume 170, 25 December 2015, Pages 251-256.

[21] A Ian McLoughlin; Haomin Zhang; Zhipeng Xie; Yan Song; Wei Xiao,"Robust Sound Event Classification Using Deep Neural Networks," IEEE/ACM Transactions on Audio, Speech, and Language Processing,vol-23,pp. 540 - 552,2015.

[22] Xian Tang,"Hybrid Hidden Markov Model and Artificial Neural Network for Automatic Speech Recognition," Circuits, Communications and Systems, IEEE 2009.

[23] D. F. Wolf, R. A. F. Romero, E. Marques,.J. Padhye, V. Firoiu, and D. Towsley, "Using Embedded Processors in Hardware Models of Artificial Neural Networks," ResearchGate article- 2001.

[24] G.Rodrigues Hoelzle and F.Morgado Dias,"Hardware Implementation Of An Artificial Neural Network With An Embedded Microprocessor In A Fpga", centro de ciencias Matematicas, campus universitario de Penteada, Madeira, Portugal, 2009.

[25] Shreyas Patel, John Sahaya Rani Alex, Nithya Venkatesan,"Low-Power Multi-Layer Perceptron Neural Network Architecture for Speech Recognition Networks", Indian Journal of Science and Technology,Vol 8(20), August 2015.

[26] John Sahaya Rani Alex, Ajinkya Sunil Mukhedkar, Nithya Venkatesan, "Performance Analysis of SOFM based Reduced Complexity Feature Extraction Methods with back Propagation Neural Network for Multilingual Digit Recognition Networks", Indian Journal of Science and Technology, Vol 8(19), IPL098, August 2015.

[27] Alvarez, A. G., Evin, D. A., & Verrastro, S. (2016), ''Implementation of a Speech Recognition System in a DSC,''IEEE Latin America Transactions, 14(6), 2657–2662.

[28] U. Suryawanshi and S. R. Ganorkar, "Hardware Implementationof Speech Recognition Using MFCC and Euclidean Distance," Int. J. Adv. Res. Electr. Electron. Instrum. Eng., vol. 03, no. 08, pp. 11248–11254, Aug. 2014.

[29] S. Li and H. Ren, "An isolated word recognition system based on DSP and improved dynamic time warping algorithm," IEEE Int. Conf. Prog. Informatics Comput, vol. 1, pp. 136–139, 2010.

[30] K. Joshi,N. Kolhare, and V. M. Pandharipande, "Implementation of Speech Recognition System using DSP

Processor ADSP2181," Int. J. Electron. Signals Syst., vol. 1, no. 3, 2012.

[31] J. XinXing and S. Xu, "Speech Recognition Based on Efficient DTW Algorithm and Its DSP Implementation," Procedia Eng., vol. 29, pp. 832–836, Jan. 2012.

[32] Y. Meng, "Speech Recognition on DSP: Algorithm Optimization and Performance Analysis," Ph.D. dissertation, University of Hong Kong, 2004.

[33] T. Sledevic,G. Tamulevicius, and D. Navakauskas, "Upgrading FPGA Implementation of Isolated Word Recognition System for a Real- Time Operation," Elektron. ir Elektrotechnika, pp. 123–128, 2013.

[34] A. Aldahoud,H. Atoui, and M. Fezari, "Robust Automatic Speech recognition System Implemented in a Hybrid Design DSP-FPGA," Int. J. Signal Process. Image Process. Pattern Recognit., vol. 6, no. 5, pp. 333–342, Oct. 2013.

[35] S. Pan, C. Lai, and B. Tsai, "The implementation of speech recognition systems on FPGA-based embedded systems with SoC architecture," Int. J. Innov. Comput. Inf. Control, vol. 7, no. 11, pp. 6161–6175, 2011.

[36] M. Dharsmale and M. Mahamune, "Robotic Automationthrough Speech Recognition," Int. J. Sci.Res. Publ., vol. 3, no. 6, pp. 1–4, 2013.

[37] H. Heidari,S. Gobee, and N. Jaiswal, "Isolated Word Command Recognition for Robot Navigation," Eng. Procedia, vol. 41, no. IRIS, pp. 412–419, Jan. 2012.

[38] A. Vijayaraj and N. Velmurugan, "Limited speech recognition for controlling movement of mobile robot implemented on atmega162 microcontroller," Int. J. Eng. Sci. Technol., vol. 2, no. 10, pp. 347– 350, 2009.

[39] N. Kandpal, Y. Mandke, and A. Patwardhan, "Implementation of Voice Recognition in Low Power Microcontroller," in Int. Proc. Comput. Sci. Inf. Technol., vol. 30, 2012, pp. 111 115.

[40] C.-H. Chang,Z.-H. Zhou, S.-H. Lin, J.-C. Wang, and J.-F. Wang, "Intelligent appliance control using a low-cost embedded speech recognizer," in Int. Conf. Comput. Netw. Technol., no. 1, 2012, pp. 311– 314.

[41] Q. Qu and L. Li,"Recognition Module Based on STM32," in Int. Symp. Commun. Inf. Technol., 2011, pp. 73–77.

[42] B. Kamdar, M. Bhisham, and D. Shah, "Real Time Speech Recognition using IIR Digital Filters Implemented on an Embedded System," in Int. Conf. Commun. Inf. Comput. Technol, 2012, pp. 1–5.

[43] V. Naresh,B. Venkataramani, A. Karan, and J. Manikandan, "PSoC based isolated speech recognition system," in Int.Conf. Commun. Signal Process, 2013, pp. 693–697.

[44] Y. Xing and W. Chen,"Design of speech recognition robot based on MCU," in Int. Conf. Intell. Human-Machine Syst. Cybern., vol. 1. Ieee, Aug. 2012, pp. 253–256.

[45] H. Liu, Y. Qian, and J. Liu, "English speech recognition system on chip," Tsinghua Sci. Technol., vol. 16, no. 1, pp. 95–99, 2011.

[46] S. K. Nanda and A. P. Dhande, "Microcontroller implementation of a voice command recognition system for human-machine interface in embedded systems," Int. J. Electron. Commun. Soft Comput. Sci. Eng.,vol. 1, no. 1, pp. 5–8, 2005.

[47] Alhanjouri, M., Lubbad, M. A., & Alkurdi, M. Z. (2013),"Robust Speaker Identification using Denoised Wave Atom and GMM," International Journal of Computer Applications, 67(5), 2013.

[48] Mark, P.W., G.S. Rash, P.M. Quesada, A.H. Desoky, "Wavelet based Noise Removal for Biomechanical Signals: A Comparative Study", IEEE Transactions On Biomedical Engineering, 47 (2), 360–360, 2000.

[49] Nitin Trivedi, Dr. Vikesh Kumar, Saurabh Singh, Sachin Ahuja, Raman Chadha, 2011,"Speech Recognition by Wavelet Analysis," International Journal of Computer Applications (0975 – 8887) Volume 15– No.8.

[50] John, P., Mahesh, T. Y., & Sebastian, B. (2017),"ECG signal de-noising, optimization and classification by wave atom transform," 2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT).

[51] SANJAYA, Mada et SALLEH, Zabidin. Implementasi Pengenalan Pola Suara Menggunakan," Mel-Frequency Cepstrum Coefficients (MFCC) dan Adaptive Neuro-Fuzzy Inferense System (ANFIS), '' sebagai Kontrol Lampu Otomatis. ALHAZEN, 2014, vol. 1, no 1, p. 43-54.

[52] A. Joshi,R. Kaur,''A Study of speech emotion recognition methods,'' International, Journal of Computer Science and Mobile Computing. 2 (2013) 28-31.

[53] A. Hassan and R. I. Damper, ''Multi-class and hierarchical SVMs for emotion recognition,'' In Proc. Interspeech, 2010 .

[54] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," ACM Trans. Intell. Syst. Technol., vol. 2, no. 3, pp. 1–27, Apr. 2011.

[55] H.Gyulyustan, S.Enkov, "Experimental speech recognition system based on Raspberry Pi 3", IOSR Journal of Computer Engineering (IOSR-JCE), Volume 19, Issue 3, PP 107-112, May.-June, 2017.

[56] S. Attawibulkul, B. Kaewkamnerdpong, Y.Miyanaga, ''Noisy Speech Training in MFCC-based Speech Recognition with Noise Suppression Toward Robot Assisted Autism therapy,'' The 2017 Biomedical Engineering International Conference (BMEiCON-2017).

[57] S.Ahmed Rahat, A. Imteaj and T. Rahman,'' An IoT based Interactive Speech Recognizable Robot with Distance control using Raspberry Pi,'' 2018 2nd Int. Conf. on Innovations in Science, Engineering and Technology (ICISET), 27-28 October 2018, Chittagong, Bangladesh.

[58] Ssarma, M. K., Gajurel, A., Pokhrel, A., & Joshi, B. (2017). ''HMM based isolated word Nepali speech recognition,'' 2017 International Conference on Machine Learning and Cybernetics (ICMLC).

[59] S.Prasad Nandyala, T.Kishore Kumar,'' Real Time Isolated Word Speech Recognition System for Human Computer Interaction,'' International Journal of Computer Applications (0975 – 8887), Volume 12– No.2, November 2010.

[60] T.Sledevi, D.Navakauskas,''FPGA Based Fast Lithuanian Isolated Word Recognition System,'' Euro-Conference, Zagreb, Croatia, 1-4 July 2013.