IRAQI Academic Scientific Journals
Al-Rafidain Engineering Journal (AREJ)
Vol. 24, No 2, December 2019, pp. 23-35
ISSN (Print) 1813-0526;
ISSN (Online) 2220-1270

Real-Time Voice Transmission over Wireless Sensor Network (VoWSN) based Automatic Speech Recognition (ASR) Technique

Ina'am Fathi * inamfth@gmail.com
Qutaiba Ibrahim Ali** qut1974@gmail.com
Jassim M. Abdul-Jabbar*** drjssm@gmail.com
Departement of Computer Engineering,Collage of Engineering, University of Mosul

Received: 2019-12-28 Accepted: 2020-01-20

ABSTRACT

The speech recognition process under the embedded system with constrained resources represents a challenge in terms of processing capability, storage memory, and B.W (or data rate). So, in this paper an efficient Real-Time Voice over Wireless sensor network (VoWSN) platform based on Automatic Speech Recognition (ASR) system to be used in emergency scenarios is proposed, implemented and evaluated. The workflow principle of the proposed system is depending on a Category Transformation Protocol (CTP) that transforms system category gradually from network dependent ASR system with a full dictionary and language model (i.e. large vocabulary) to fully embedded ASR system with customized dictionary and language model (i.e. small vocabulary). Moreover, a comparison study has been performed between our proposed VoWSN based ASR system and a VoWSN based streaming system. This comparison is performed to elaborate the gains achieved when sending the text of the voice signal instead of sending the voice signal. Additionally, the Voice over IoTs (VoIoTs) system has been evaluated utilizing Voice streaming or ASR system to evaluate the system performance when connecting to the Internet. The comparison evaluation process is achieved by means of experimental platform and simulation.

Keywords:

VoWSN, ASR system, Streaming system, VoIoTs, embedded systems.

https://rengj.mosuljournals.com
Email: alrafidain_engjournal@umosul.edu.iq

==========

1. INTRODUCTION

Wireless voice transmission represents an expeditious communication mechanism and therefore unsurprising that it is always wanted in a wide range of emergency scenarios [1]. Generally, in emergency scenarios the time plays a dominant role in the rescue operation. For an efficient communication mechanism, voice can be a significant interfacing tool between the persons in the disaster zone and the rescue centre. Typically, voice interface doesn't require visual or physical contact with WSN devices. As a result this feature can speed up the input process and consequently speeding up the rescue operation. However, Real-time voice transmission applications have strict requirements by means of end-to-end delay, data losing and dedicated Bandwidth (B.W) during passing the network. So transmitting audio data from disaster zone toward a rescue centre with an acceptable quality within Real-Time constraints is an important issue.

ASR technology provides tools to transform human voice signals into text. Consequently this text can be sent toward a rescue centre with less B.W. and power compared to sending original audio stream. So in this paper we aim to evaluate the performance of a proposed VoWSN based on ASR system for efficiently voice transmission system within Real-Time constraints to be used in emergency scenarios. Also we aimed to compare the performance of the proposed

This paper is organized as follows: section 1 includes an introduction to the objectives of this research. Section 2 includes a literature review. Section 3 includes a brief background of ASR system with an explanation of system evaluation procedures. Section 4 gives an introduction to the ASR technique based embedded platforms and its categories architecture. Section 5

describes our suggested protocol and algorithm design for VoWSN based ASR system. Section 6 illustrates the implementation of suggested ASR sysem on RPi3 platform. Section 7 describes a set of suggested performance evaluation methodologies proposed for the evaluation of the suggested VoWSN. And finally section 8 includes conclusions and discussion.

2. LITERATURE REVIEW

The speech recognition process of embedded system is realized by some research, but still has a lot of aspects which needs to be improved [2]. However, a few research endeavours have integrated speech recognition technology into WSN systems and explored general framework practicality. The earliest attempt of implementing speech recognition system on embedded resource constrains systems introduced by S. Phadke, et al.. They combined the aspects of both hardware and software design of implementing a speaker dependent, isolated word speech recognition system. They used modified Mel-scaled Frequency Cepstral Coefficients (MFCC) as feature extraction method and Dynamic Time Warping (DTW) as template matching process [3].

Also, the authors C. Shen, et al., presented the design and implementation of a distributed sensor network application for embedded, isolated-word, real-time speech recognition system. They adopted a parameterized-dataflow-based modelling approach to model the functionalities associated with sensing and processing of acoustic data. The associated embedded software implemented on an off-the-shelf sensor node platform and a TDMA access protocol was developed to manage the wireless channel [4].

In addition, the authors F. Sutton, et al., demonstrated the implementation of prototype architecture for automatic single word speech recognition on resource-constrained embedded devices. The experiments results showed that the prototype achieved a high average detection rate of 96%, while only dissipating 28.5 mW for continuous audio sampling and duty-cycled speech recognition. Audio signal acquisition was performed by a dedicated audio codec [5].

Moreover, the authors, R. P. Raghava, et al., presented the design and implementation automatic single word speech recognition system on embedded devices. The words which are spell are stored in the operating system of Raspbian OS which is implemented on Raspberry Pi hardware kit of ARM 11 processor [6].

Also, the authors, G. G. tolya, et al., analysed ASR performance on utterances recorded by means of wireless sensors. The sound quality of utterances recorded by such sensors contrasts fundamentally from that of the larger audio data bases usually used for acoustic DNN (Deep Neural Network) training due to the small microphone installed on these devices. They could accomplish a5% improvement in terms of relative error reduction [7].

3. AUTOMATIC SPEECH RECOGNITION (ASR) BACKGROUND

Speech Recognition (is also known as Automatic Speech Recognition (ASR)) is the process of converting a speech signal to a sequence of words, by means of an algorithm implemented as a computer program [8]. Before going into details of speech recognition based WSNs subject, some aspects of ASR technique must be explained. Fig. 1 describes basic processing stages that are part of the ASR system and which will be used for clarifying the classification of ASR system based WSN [9].

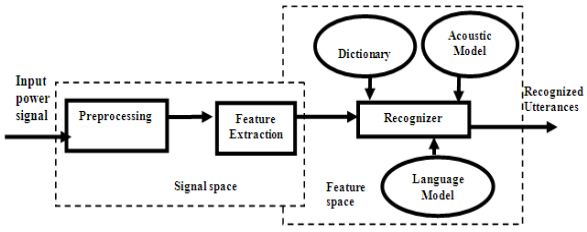

Fig. 1 Basic Architecture of ASR system.

In typical ASR systems three models are used:

- Acoustic model: is used to translate data from an audio signal to the most probable phones uttered [8]. Hidden Markov Models (HMMs) are the most powerful parametric model at the acoustic level [10]. HMM is using the forward algorithm and Viterbi algorithm [11].
- Phonetic dictionary: maps the relationship between words and its phones. This model can be

very large which greatly affects the decoding time of speech recognition.

- Language model: contains statistical information about which words should follow each other in a sentence.

As shown from Fig. 1, the speech recognition process can be divided into three consecutive processes [12]:

Pre-processing: The main tasks of pre-processing process are removing the noise contained in speech signal, filter out any parts of the speech signal that do not contain any speaking [13] and re-sample the audio signal to the correct format.

- **Feature extraction**: Feature extraction process is the most important step in speech recognition and affects greatly the performance of ASR systems. The function of feature extraction phase is extracting features from speech signal and representing them using an appropriate data model of the input signal. This process also known as Front-End analysis[14]. Mel-Frequency Cepstrum Coefficients (MFCC), Discrete Wavelet Transform (DWT), Linear Predictive Coding (LPC) are examples of feature extraction techniques those are commonly used in speech recognition [13]. MFCC has been found to be more robust in the presence of background noise compared to other algorithms [3].
- **Classification:** In ASR systems there are three approaches for classification process [13]: **Acoustic Phonetic Approach, Artificial Intelligence Approach, Pattern-Matching Approach**, **Acoustic** In **Phonetic Approach** the speech recognition relies on finding speech sounds and giving specific labels to these sounds. While **Artificial Intelligence Approach** attempts to mechanize the recognition process according to the way a person applies its intelligence in visualizing, analysing and finally making a decision on the measured acoustic features. The **Pattern-Matching Approach** has become the predominant method for most modern speech recognition systems. HMM approach is used by most systems to represent the basic units of speech. It is widely used because it is easy, simple and reliable, it can be automatically trained and feasible to use [15].

3.1. ASR performance Evaluation Criteria

The performance evaluation of a typical ASR system covers three important aspects: Speed, average Word Error Rate (WER) and Accuracy. For evaluating the accuracy, the WER must be measured firstly. The meaning of each metric is:

Speed: This parameter represents the search time on the recognizer and is measured in term of Real Time Factor (RTF) often abbreviated as xRT and calculated as the ratio between the amount of time required to decode an utterance and the length of the utterance using the Formula (1):

$$\text{xRT} = \frac{\text{Input speech recognition time}}{\text{Speech duration}} \quad (1)$$

For example, a real-time factor of 0.4 xRT means that each second of audio requires 0.4 seconds to decode (lower RTF means faster decoding) [16].

Word error rate (WER): is a metric for evaluating the accuracy of a speech recognition system. Calculating WER is a way of measuring the number of errors that occurred during the decoding of an audio signal. Generally, WER is calculated by summarizing the total number of errors in the hypothesis and dividing it by the total number of words in the correct sentence. An error is an incorrect substitution(S) or deletion (D) or insertion (I) of a word which differs the hypothesis from the correct sentence. With this, the word error rate can be calculated using Formula (2) [16]:

$$\text{WER} = \frac{(\text{I} + \text{D} + \text{S})}{\text{N}} \quad (2)$$

Where: **I**: Are the insertion words, **D**: Are the deletion words, **S**: Are the substitution words, **N**: Are the total number of words. The WER is usually measured in percent. Smaller WER means more efficient ASR system.

Accuracy: It is almost the same as the WER, but it doesn't take insertions into account as described in Formula (3).

$$\text{Accuracy} = \frac{(\text{N} - \text{D} - \text{S})}{\text{N}} \quad (3)$$

4. DESIGN AND IMPLEMENTATION OF AN ASR SYSTEM BASED EMBEDDED PLATFORMS

Generally, integrating speech recognition into a WSN system produced a new type of applications that have distributed configurations such as smart home automation, health care system, security and so on. Consequently, this research domain has been attract considerable attention for creating innovative energy -saving architectures, algorithms, and protocols to overcome constraints of these embedded systems and meeting the requirements of these applications [17]-[18]. Although many audio monitoring systems utilizing wireless sensor networks are developed over the last few years, but few of these systems are especially targeted at human speech recognition as the main goal of their designs [19]. Fig. 3 shows the basic structure of audio wireless sensor network monitoring system.

Fig. 3 Basic Structure of Audio WSN.

Typically, speech recognition is a high demanding process in terms of computational power and memory storage. Therefore, ASR based embedded systems need to overcome the capability limitations of WSN devices that defined by:

- Limited computational capability.
- Limited storage space.
- Limited bandwidth.
- Limited energy supply.

As a strategy to resolve the above limitations, a careful compromising between the computation operations and the communication tasks need to be done. This is often achieved by splitting the different functionalities of ASR process between WSN nodes and the Server. The next paragraph discusses strategies for implementing the different categories of ASR system based WSN.

4.1. Categories of ASR Based WSN System

Depending on the decision of where placing the ASR components, three categories of ASR systems based embedded devices can be recognized [9]:

a) Network Speech Recognition (NSR) system.
b) Distributed Speech Recognition (DSR) system.
c) Embedded Speech Recognition (ESR) system.

Each category adopts the client-server architecture and distributes ASR functionalities between the WSN nodes and remote server driven by many factors including device and network resources, ASR components complexity and application demanding.

NSR system is known to be server-based due to the fact that the overall ASR process is performed by the remote server and the WSN nodes are responsible of capturing, pre-processing and transmitting the speech signal to the server. Fig. 4 shows the basic operations distribution between the WSN node and remote server for NSR system. Google APIs, Alexa are approaches for NSR system. The NSR architecture has low requirement for the client devices (i.e. WSN devices).

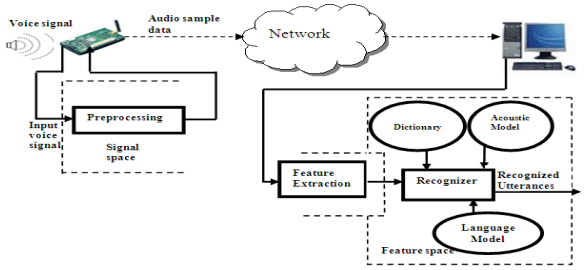

Fig. 4 Basic operations distribution in NSR system.

However, NSR suffers from network dependency and recognition degradation due to using low bit-rate speech codec. Moreover, for Real-Time application, NSR is not an optimal option due to the different sourced of delay which introduced from compression/decompression delay, network delay and recognition delay.

DSR system is known to be Client-Server ASR system where the WSN node is responsible of capturing the speech signal, pre-processing it, extracts the acoustical features from the voice and sends them to the Server [14]. When WSN nodes transmit the extracted voice feature values to the Server, further processing will be performed on the Server, including training, and classifying speech features. Fig. 5 shows Basic operations splitting in the DSR system.

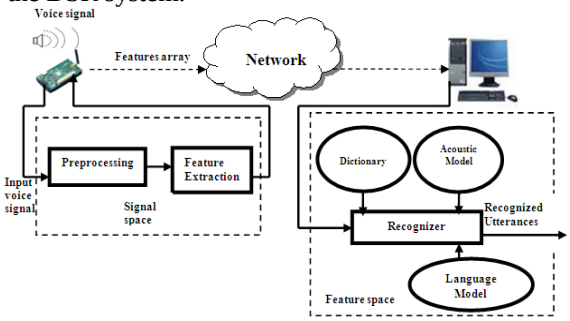

Fig. 5 Operations Distribution in *DSR* system.

In DSR system low B.W. is required compared to NSR system since the features vector (or compressed version of features vector) is only needed to be transmitted to the server. But it also suffers from Network dependency.

The ESR system is known to be a Client-based ASR system where all ASR processes are performed by the WSN node. Pocketsphinx ASR system is an approach for ESR system. Fig. 6 shows basic operations splitting in ESR system. The main advantages of the ESR architecture are:

- No communication between the remote server and the client is required. So, the ASR system does not rely on the quality of the data transmission and always ready for use.
- For Real-Time application, ESR can be a significant solution due to the elimination of coding/decoding and network delay.
- Since communication consumes power more than other operations [20], ESR is assumed to be more powerful than NSR and DSR in term of reduced power consumption.
- Low B.W. requirement compared to NSR and DSR systems since the text of recognized word is only needed to be transmitted to the server for taking actions.

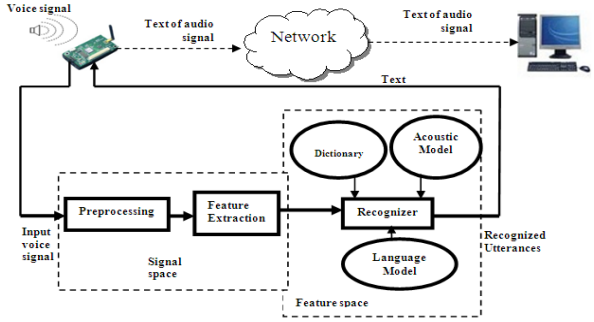

Fig. 6 Basic Operations Splitting in *ESR* system.

5. SUGGESTED ALGORITHM FOR EFFICIENT REAL-TIME VOWSN BASED ASR SYSTEM

In the past few years the foremost known architectures for ASR applications are NSR or DSR [10]. However, NSR and DSR architectures are depending on the data network and thus facing many issues in terms of: latency, losing data packets, bandwidth demanding, and security problems [10]. Recently, the vast progress of the WSN devices makes these embedded devices reach performance levels almost the same as their desktop equivalents. Thus it begins to be potential to execute the entire ASR process on the embedded devices with some restrictions and assumptions [21]. By this mean, the ASR process is known as Embedded Speech Recognition (ESR) system where the ASR process is completely performed by embedded processors.

With constrained resources of WSN devices, the process of ASR with its usual requirements faces specific challenges in hardware and algorithms design to cope the limited resources [20]. Therefore, the ASR algorithms with low computational complexity, low memory storage and low communication cost are preferred for these devices. The main feature of commercial speech recognition systems is the large size of dictionary and Language Model (LM). Generic models are very large (several gigabytes and created from large texts) and thus impractical to use them in the decoder of the recognizer [22]. However, the speed and accuracy of the recognition process is mainly affected by the size of the vocabulary. A larger dictionary means that more words need to be considered for the recognition, which means it will take longer time to map the input with the best interpretation [21]. A larger dictionary will also mean that the odds of different word combinations sounding similar increases. Additionally, larger size means more memory requirement for storage. When the dictionary size and complexity of the language model are both high, it is common to use an external server to perform the recognition at the expense of certain limitations (network availability, latency, etc). However, most recognition systems have dictionary and LM tuned to a specific domain. For example, medical dictionary describes medical dictation. This limited vocabulary may reduce the size of the dictionary and LM which has to be built [23] accordingly for that domain. So, if a system is meant to do limited speech recognition, the vocabulary can often be limited to a small set of words, and therefore the accuracy and speed can also be improved [16]. In the following paragraph, we detail the principle of an algorithm that will make ASR system running efficiently on WSN node with high accuracy.

Generally, most state-of the art ASR systems are NSR systems which relying completely on the remote server to perform the recognition process. The NSR workflow and its operation details are shown in Fig. 7 and 8 respectively.

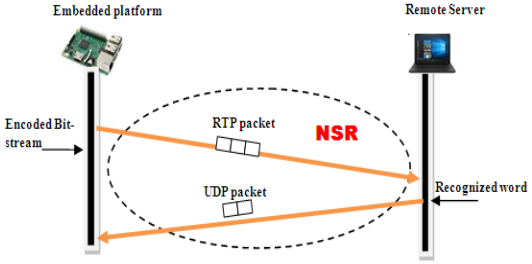

Fig. 7 NSR System Workflow.

As illustrated earlier, the principle of NSR operation depends on executing all the speech recognition operations (i.e. feature extraction and comparison) by the remote server. The duty of the embedded platform is limited only on capturing, processing and encoding voice signal with a certain encoding technique then sends the encoded word signal to the remote server as shown by steps (1), (2), (3) & (4) in Fig. 8. Consequently, the remote server will receive the encoded signal and begins to recognize it then sends back the resulting text of the recognized word to the embedded platform as shown by steps (5), (6), (7) & (8) in Fig. 8. The embedded platform will receive the recognized word and will take required actions accordingly as shown by steps (9), & (10) in Fig. 8.

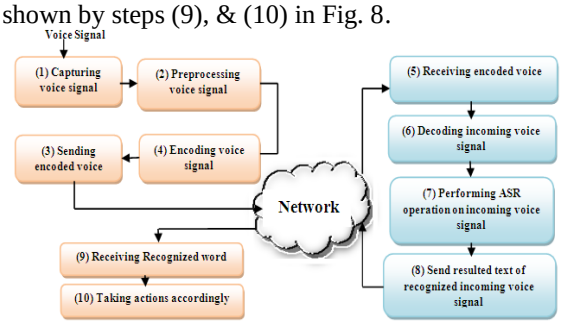

Fig.8 Details of NSR System Workflow.

Although DSR and NSR are network dependent categories and relying on the server to perform the recognition process, there are several differences between them which make the DSR more powerful and favouring than the NSR [24]:

- The speech CODECs (which are used with NSR systems) are optimized to provide the best perceptual quality while the feature extraction algorithms are optimized for giving the lowest WER.
- The ASR system needs some set of characteristic parameters and does not require the high quality speech. Therefore, it needs lower data rates.
- Because the feature extraction process is implemented by the client, so the higher sampling rates which cover full bandwidth of the speech signal are possible.
- DSR systems are not constrained to the error-mitigation algorithm of the speech codec, so it can

develop better error-handling procedures in terms of WER.

The principle of our suggested algorithm for VoWSN based ASR system is based on transforming the system category from network dependent system, i.e. NSR system to partially DSR system to reach finally completely network independent system, i.e. ESR system. The gradually category transformation is performing by a suggested Category Transformation Protocol (CTP). For simplicity we assumed the remote server as the destination of our system. Therefore we called it the Remote and Destination (RaD) server. The workflow details of CTP between the embedded system (i.e. WSN) and RaD server (which is a laptop in our work) are shown in Fig. 9. For the above listed reasons, we based on the DSR system as a primary ASR system when there is neither dictionary nor LM at WSN node.

Before starting up suggested algorithm actions, system initialization has to be performed. The initialization process is represented by installing ASR system on the WSN node and the RaD server in the following manner:

- ASR system for recognition has to be installed on the WSN node with an empty Dic. and LM.
- ASR system for recognition has to be installed on the RaD server with a default Dic. and LM.

As shown in Fig. 10, the algorithm suggests that: At system start-up, the Dic. and LM are empties and the WSN node does not have the ability to perform the recognition process therefore its duty is limited on capturing, pre-processing and extracting features of first incoming voice signals. The features vector is then sent to the RaD server for comparison and text extraction. Consequently, the remote server receives features vector and begin to recognize it with a default dictionary and LM then send back the resulting text of the recognized word to the WSN node

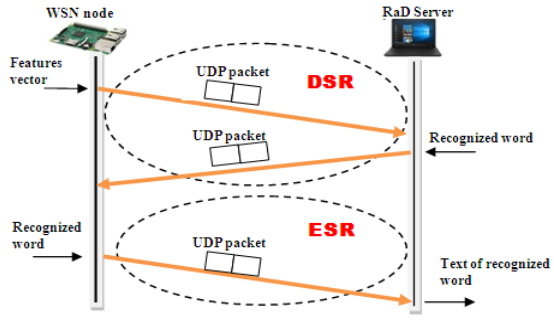

Fig. 9 CTP Workflow

- On receiving the text, WSN node will create a corpus file and saving the receiving text in it then create dictionary and language model files accordingly.
- Once the Dic. and LM are created, the system is able to execute the search process locally.
- Subsequently, for each incoming new word signal, ASR engine at WSN node will do the following operations:
- Trying to extract the text of this word if its text available then sent the recognized word text to the RaD server for taking actions accordingly.
- While if the text of the new word is unavailable then the features vector of the new word sent to the RaD Server for performing recognition process. At this point, the system acts as DSR system where only features vector are sent to the Remote Server for text extraction.
- So, the dictionary and language models will continually updated for each new incoming word. This updating operation continues till reaching optimized Dic. and LM.
- Gradually, the dictionary and LM will stripped down and the system will reach a point when there is no need to contact the RaD Server for text extraction.
- Finally, the Dic. and LM will be optimized with a size much smaller than the default. At this point the extraction of the text for each incoming word is done WSN node, and only the text of the recognized word will be sent to the RaD Server. At reaching this state, the system will act as ESR where all the operation of capturing, pre-processing, feature extraction and comparison is done locally at WSN node.
- Now the system is ESR system and only the text of each incoming word will be sent to the RaD Server for taking actions accordingly.

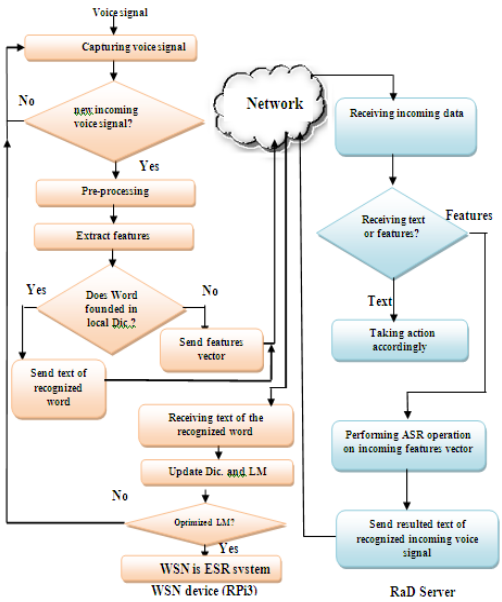

Fig. 10 Suggested Algorithm Flowchart.

6. VoWSN BASED ASR SYSTEM HARDWARE AND SOFTWARE IMPLEMENTATION

The Main components of our proposed VoWSN based ASR system is shown in Fig. 11. The system is composed of WSN device which is RPi3 model B equipped with a USB sound card associates with a microphone, Wi-Fi router and laptop which is assumed to be a RaD Server. As explained in Fig. 11, the functionality of WSN device is to perform ASR process on each incoming voice signal and sending the resulting associate text to the RaD Server. This system is designed to act as ESR system with optimized Dic. and LM model.

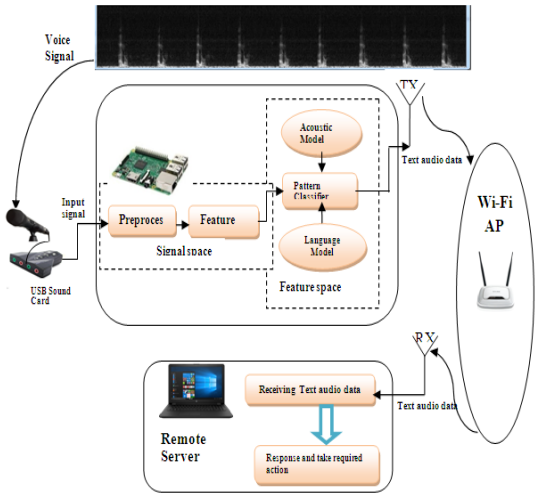

Fig.11 Proposed VoWSN based ASR system.

However, selecting the ASR system that is to be implemented on our suggested VoWSN based ASR system has to take into account the limited resources of embedded platform for efficient performance. Nowadays, there are several ASR products available, from commercial ones Microsoft Windows Speech Recognizer (WSR), to open source software packages like Sphinx or Hidden Markov Tool Kit (HTK) [25]. For implementing our work, a lightweight ASR project PocketSphinx is chosen for this purpose due to its many valuable features.

6.1. Pocketsphinx Structure Overview

PocketSphinx is one of CMU Sphinx projects developed at Carnegie Mellon University (CMU). PocketSphinx is lightweight, free, open source, Internet independent, large vocabulary and speaker-independent continuous speech recognition engine dedicated for real-time applications in embedded systems [26]. Pocketsphinx uses a probabilistic approach while doing the translations and uses HMMs to translate speech into text. But to ensure satisfactory performance regarded to the limitations of embedded devices, PocketSphinx uses less accurate Gaussian Mixture Model (GMM) [26].

The main features of Pocketsphinx are its vocabulary configurability and offline working. The vocabulary configurability enables a user to create models on demand. While the offline working means the ability of performing the ASR process locally and there is no need to send user speech through the network to the remote server and include strange items in user agreement. So, when it needed a system which recognizes a small set of commands without the internet, PocketSphinx may be a preferred solution.

According to the Pocketsphinx structure, it combines three models in to do the match [26]. These models are:

- Acoustic model: There are a number of acoustic models trained for different languages such as US-English model.
- Phonetic dictionary: Maps phones to words. In PocketSphinx, *.dic* files are dictionary files.
- Language model: The most commonly used language models are N-gram language models. These models assume that the probability of any word in a sequence of words depends only on the previous N words in the sequence. PocketSphinx uses tri-gram language (where n=3) model and the *.lm* files are language model files.

The Dic. and LM will be generated specifically for the implementation of the proposed ASR system, while the acoustic model will be a PTM acoustic model of Pocketsphinx. This acoustic model significantly constrains the number of Gaussians in each mixture to improve performance and dedicated for embedded platforms

For the language model, a statistical language modelling approach is used. To create statistical English-language models, various software packages have been developed by Speech Group at CMU and in use for many years such as *lmtool* [27]. For this research, we used *lmtool* to generate our customized English language model.

.

7. EVALUATION PROCESS OF VOWSN BASED ASR SYSTEM

The objective of this research is to elaborate the gains those achieved by implementing our proposed VoWSN based ASR system compared to VoWSN based streaming system through the evaluation process. The final goal is to aid the developer to select an efficient technique in a particular situation that requires Real-Time VoWSN. The Evaluation Process composes of two phases:

1. Definition of the evaluation process methodology.
2. Implementation of the evaluation process methodology.

7.1. Definition of the Evaluation Process Methodology

Firstly, we aimed to explain the methodology that will be followed for evaluation process. As shown in Fig. 14, the three stages of our evaluation methodology are:

1. Definition of evaluation strategy.

2. Selection of evaluation scenarios.
3. Selection of evaluation criterion.

7.1.1 Definition of Evaluation Strategy

The strategy that we adopted depends on highlighting the benefits of our suggested system and feasibility of implementing it in scenarios those requires Real-Time voice transmission. The workflow strategy that shown in Fig. 14 is composed of the following steps:

1. Firstly, the performance of the proposed system is evaluated practically by implementing it on RPi3 device.
2. Secondly, depending on the results obtained in 1, the system is modelled by means of simulation.
3. Then, the alternative solution, VoWSN based streaming system, is implemented on the same platform and evaluated using two different audio encoding techniques.
4. Depending on the results obtained in 3, the system is modelled by means of simulation for more analysis.
5. Finally, after that the two systems are implemented but now connecting to the Internet to perform VoIoTs system and their performance is compared by means of simulation.

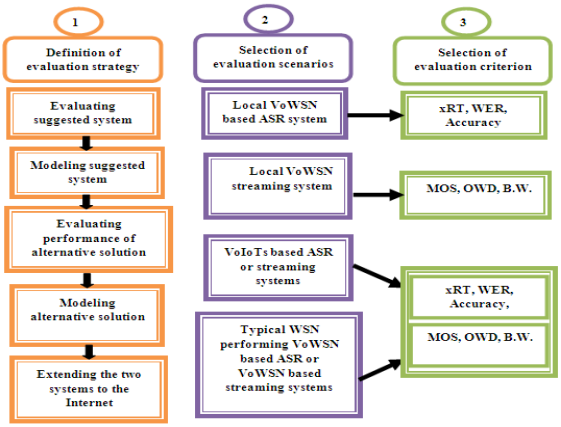

Fig.14 Stages of Evaluation process methodology.

7.1.2. Selection of Evaluation Scenarios

The main step in the evaluation methodology is to define the scenarios that will be used for the evaluation process. In order to obtain an extensive evaluation of the suggested VoWSN based ASR system; four scenarios are defined for performance comparison as follows:

a. Local ASR System (i.e. without Internet).
b. Local Voice Streaming (i.e. without Internet).
c. Voice over IoTs (VoIoTs) utilizing Voice streaming or ASR system.

These four scenarios are deployed respectively from 1 to 4 to build an overview that will help us to clarify the gains obtained from applying the proposed VoWSN based ASR algorithm in term of maximum number of WSN nodes those the system can accommodate within Real-Time constraints.

7.1.3. Selection of Evaluation Metrics

The evaluation of our suggested ASR system will cover the three aspects:

- Decoding Speed.
- WER.
- Accuracy.

While Streaming based WSN system will be evaluated by measuring three important parameters:

- Mean Opinion Score (MOS).
- One-Way Delay (OWD).
- Bandwidth (B.W.) (or throughput).

OWD and B.W. are calculated using Formulas (4) & (5) respectively as illustrated in [28]:

$$\text{one} - \text{way delay}(P_{i+n}) = R_{i+n} - S_{i+n} \ [S] \quad (4)$$

B.W = total packet size (bits)*[(codec bit rate)/ Total payload size(bits)]=B.W bps (5)

7.2. Implementation of the Evaluation Process

The implementation of the evaluation process includes the realization of proposed scenarios those defined by the process evaluation methodology. These scenarios are implemented by means of practice and simulation.

7.2.1 Local ASR System Scenario

In this scenario we aimed to evaluate the proposed ASR system with Real-Time limits which are represented by RTF<1. However, trade-off between accuracy and speed is depending on the application demanding. In Real-Time applications such as emergency scenarios, the speed is playing an important role in rescue operation. It is relatively easy to improve recognition speed while trading away some accuracy, for example by reducing the search space, and by using simpler acoustic and language models [29]. Our goal is to improve the accuracy of the ASR system within the Real-Time limits. This scenario is deployed in two steps:

- By means of practice: where the proposed ASR system is implemented on the proposed ESR platform that explained in section V.
- By means of simulation: where the proposed ASR system is modelled using OPNET simulation tool.

7.2.1.1 Practice implementation

For practical implementation, baseline PocketSphinx is used as the ASR engine. PocketSphinx can use different types of input:

- Input based on a single file: An audio file in .wav format is read by PocketSphinx and used in the speech recognition process.
- Continuous input-stream: Continuous audio from a microphone of an audio-stream is used.

For our practice testing, we used a standard audio file from Harvard which was downloaded from [30]. Then the file is converted to the format that is required by Pocketsphinx tool which is: Sampling rate=16 kHz, Number of bits=16bit Little-Endian, Number of channels=Mono (single channel). The file contains 10 phrases with 76 words. The PocketSphinx is operated on RPi3 (with 1.2GHz single core processing) in two manners:

- Utilizing the optimized vocabulary (small Dic. and LM) which is containing 10 in-model phrases with 76 words.
- Utilizing default vocabulary that associates with PocketSphinx package (Full Dic. and LM).

However, the default settings of Pocketsphinx are not enough to achieve Real-Time performance (i.e. RTF<1) on most tasks. For efficient performance, some command-line flags should be experimented. The main parameters those are needed to configure search width and thus accuracy-performance balance and their meanings are shown in Table 2 [31]:

Table 2 Pocketsphinx parameters settings.

Parameters	value	Meanings
Ds	2	Frame GMM computation down-sampling ratio
Topn	4	Maximum number of top Gaussians to use in scoring
Maxwpf	5	Maximum number of distinct word exits at each frame
maxhmmpf	3000	Maximum number of active HMMs to maintain at each frame
Pl_window	10	Phoneme lookahead window size, in frames

After configuring the decoder with the required parameters those explained in Table 2, Pocketsphinx now ready to be executed and the performance parameters can be measured using `word_align.pl` tool. `word_align.pl` tool is a script supported by Pocketsphinx and it is a part of Sphinxtrain distribution. The measuring process is performed in the following sequence:

1. Dumping speech utterances (speech to be recognized) into .wav files.
2. Writing the reference text files associate with the speech utterances.
3. Using decoder to decode it by running `pocketsphinx_batch` script.
4. Finally, running the test by using `word_align.pl` script.

The script `word-align.pl` will report the WER, accuracy and decoding speed which is represented in term of xRT. Table 3 shows the results of practical implementation of ASR system using optimized dictionary and LM.

Table 3. ASR Practice implementation Results.

Acoustic model	Memory usage%	WER%	accuracy%	xRT	Total time(sec)
Opt. dic.	6.3	9.21	90.79	0.327	9.4
Full dic.	14.5	76.32	23.68	4.288	120.71

The speed of decoding (xRT) is calculating using the Formula (5). Using Pocketsphinx, Formula (5) is translated by means of the 3 stages of Viterbi search algorithm. So the total xRT is estimated by adding the three stages xRT: TOTAL fwdtree xRT, TOTAL fwdflat xRT and TOTAL bestpath xRT where:

- fwdtree : is a Forward Tree Viterbi beam search=0.213
- fwdflat: is a Forward Flat Viterbi beam search=0.112
- bestpath: is an N-best list search of alternative hypotheses=0.002

So, Decoding Speed = 0.213 + 0.112 + 0.002 = 0.327xRT which means 1 second of speech is decode in 0.327 seconds of CPU time with optimized Dic. and LM. While the speed of decoding with Full Dic. and LM is equal to 4.288xRT which means 1 second of speech is decode in 4.288 seconds of CPU time. However, WER can be enhanced greatly by using another version of acoustic model, but it costs of RTF to be > 1. This enhancement in accuracy is required in such application despite Real-Time limits.

As seen from Table 3, the recognition with optimized Dic. gets a WER reduction of 67.11% versus the WER obtained with full Dic. This enhancement is greater than the enhancement obtained by [23] which achieved a reduction in WER about 20.4% by modifying the LM and Dic. to closer resemble typical TED-talk contents but the cost of RTF which passed the 1.0 limit. Also, our WER reduction is much greater than WER reduction obtained by [21] which depended on Language Modelling Switching (LMS) mechanism for topic and LM optimization. They used three sets of sentences and each set composed utterances belonging to different domain. Additionally, the results show that the RTF (xRT) is enhanced by 396% with respect to the RTF obtained when recognizing using full Dic.. Also this enhancement is much greater than RTF obtained by [21].

7.2.1.2 Simulation Implementation

The purpose of the simulation is to examine the feasibility of utilizing our proposed ASR system in real situation. This is reflected in the maximum number of WSN nods running ASR application those the system can accommodate within Real-Time constraints.

There are several traditional network simulators used by many researchers for WSN simulation. The most common one is OPNET package [18]. We used OPNET 14.5 to deploy our work. As shown in Fig. 15, the simulation network model for VoWSN based ASR system consists of WSN node, Server and Access Point (AP) as the main components of our system. Also the network model contains application configuration and profile configuration for specifying and configuring our ASR application.

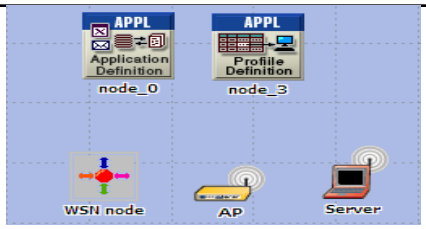

Fig. 15 Network Model for ASR Scenario.

For simulation simplicity and accurate results some assumptions are made. The assumptions are:

1. WSN nodes are fixed.
2. Each WSN Node processing speed is assumed to be equal to 20000packet/sec to simulate the processor speed of RPi3 which is 1.2 GHz. This speed value is calculated according to [32].
3. Remote Server speed is assumed to be infinity.
4. For simulation validation, an application that simulates the ASR application is proposed and implemented according to the results obtained in practical implementation as follows:

As shown from Table 3, ASR based optimized Dic. and LM takes about 9.4s to recognize a whole file which containing 76 words and this means that one word recognizing delay ≈ 0.123s. So, the average processing time (translation time) for each word is equal to 0.123 sec. This means in 1 sec about 8 incoming spoken words can be recognized and sent in one packet to the server. The length of each word is assumed to be 5 bytes according to [33]. So, each packet contains 8 words each of length 5 bytes.

5. There is no other network traffic besides ASR traffic.

The results of this simulation are shown in Table 4.

Table 4 ASR Simulation Results.

802.11g speed	Max WSN nodes	Throughput (kbps)	Packet-end-to-end-delay (s)
1M	59	125	0.155
11M	181	385	0.136
24M	335	710	0.136
54M	352	747	0.132

The maximum numbers of WSN those obtained in each result are depending on voice streaming Real-Time constraints. These two important constraints are:

- One-way delay (Packet-end-to-end-delay): One-way delay must be < 150 [34].
- Packets loss: is a critical specification because one lost packet means 8 lost words. So this parameter should be = zero for our suggested ASR system.

7.2.2 Local Voice Streaming Scenario

In this scenario we test the performance of the alternative solution for the VoWSN which is the voice streaming practically then validate the results by means of simulation. For this scenario, the voice streaming process is performed using G.711 encoding algorithm. Generally, G.711 algorithm supposed to be high data rate encoding technique with a data rate = 64 kbps (i.e. uncompressed data). Then the system is simulated using G.729 algorithm to investigate the impact of low data rate on voice streaming with a data rate equal to 8 kbps. Finally a simulation comparison has been done between the results of G.729 encoding technique with those obtained by G.711 encoding technique. The algorithms specifications of G.711 and G.729 are listed in Table 5 [35].

Table 5 G.711 and G.729 Specifications.

Technique	Algorithm	Frame Interarrival time(ms)	Frame size (byte)	Packet Rate (Packet/second)	IP Packet size (byte)	Codec bit rate (kbps)
G.711	pulse-code modulation (PCM)	10	80	100	120	64
G.729	Conjugate Structure– Algebraic Code Excited Linear Prediction (ACELP)	10	10	100	50	8

The simulation assumptions and settings are the same as those described above except that the application is different. This study assumes there is only one way voice transmission application and there is no voice conferencing. So, a voice streaming application is proposed and implemented throughout this simulation. The results for voice streaming of practice and simulation tests utilizing G.711 encoding algorithm are listed in Table 6.

Table 6 Practice and Simulation Results for G.711.

Tools	Mean Opinion Score(MOS)	One-way Delay (ms)	B.W(Throughput) (kbps)
G.711 Practical	-	26	96
G.711 Simulation	4.1	25	96

For practical implementation of voice streaming using G.711 encoding technique, FFmpeg tool is used and operated on RPi3 in the same manner that illustrated in [28]. Also Formulas (10) & (11) are used for calculating OWD and B.W. respectively.

In the simulation implementation, a voice streaming application is designed and implemented using G.711 and G.729 algorithms with the same specifications listed in Table 5. The aim of this simulation is to find the maximum number of WSN nodes those the system can accommodate using IEEE 802.11 WLAN standard with speeds of: 1M, 11M, 24M and 54M. The results are shown in Tables 7 and 8 respectively:

Table 7 G.711 Simulation Results.

802.11g speed	Max WSN nodes	Throughput (kbps)	MOS	Packet-end-delay (s)
1M	2	192	4.1	0.0251
11M	10	960	4.1	0.0266
24M	17	1632	4.1	0.027
54M	19	1824	4.1	0.0284

Table 8 G.729 Simulation Results.

802.11g speed	Max WSN nodes	Throughput (kbps)	MOS	Packet-end-delay (s)
1M	4	160	3.6	0.027
11M	10	400	3.6	0.028
24M	19	760	3.6	0.028
54M	20	800	3.6	0.028

The maximum numbers of WSN those obtained in each result are depending on two factors:

- MOS score: for streaming voice the cut-off MOS score that can be tolerated is around 2.5 [36].
- One-way Delay (OWD): The Real-time constraints for OWD must be < 150 ms.

The International Telecommunication Union – Telecommunication (ITU-T) gives the guidelines of the delay for different types of voice quality, as shown in Table 9.

Table 9 ITU-T Precept for Voice Quality.

Delay (ms)	Voice Quality
0-150	Good
150-300	Acceptable
>300	Poor

7.2.3. Voice over IoTs (VoIoTs) Utilizing Voice Streaming and ASR system Simulation Scenario

The goal of this scenario is to study the behaviour of our proposed VoWSN when connecting to the Internet. The IEEE 802.11 WLAN standard with speeds of 54M is used. The encoding algorithms G.711 and G.729 with the specifications listed in Table 5 are used as encoding techniques. Fig. 16 shows the simulation environment of VoIoTs system.

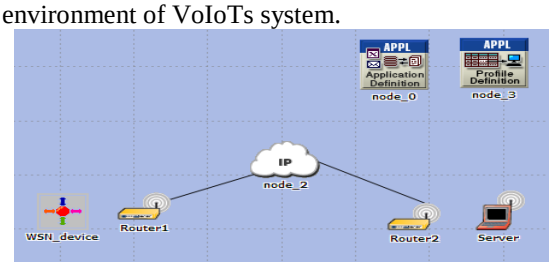

Fig. 16. Simulation Environment of VoIoTs.

Before deploying VoIoTs, the relationship between Internet delay and MOS has to be founded. Table 10 shows the relationship between Internet Delay and MOS for G.711 and G.729 algorithms.

Table10 Internet Delay & MOS Relationship.

Internet delay(ms)	G.711 MOS	G.729 MOS
0	4.1	3.6
50	3.86	3.28
100	3.56	2.9
150	3.22	2.58
200	2.88	2.23
250	2.53	1.9
300	2.18	1.6

The results show that at Internet delay 200ms and 300ms the MOS of G.729 and G.711 respectively are breakdown. So, we chose a midpoint where MOS at acceptable value and this value is obtained when the Internet delay=150 msec. This value is chosen since the cut-off MOS score for streaming voice that can be tolerated is around 2.5 [36]**.** The results of this simulation are listed in Table 11.

Table 11 VoIoTs simulation Results.

Encoding technique	Max no. nodes	Throughput (Mbps)	Packet-end_ delay (s)	MOS
ASR	424	1.8	0.275	-
G.711	15	2.82	0.176	3.22
G.729	32	2.5	0.186	2.5

8. CONCLUSIONS

The main objective of this paper is to enhance Real-Time VoWSN system through deploying a mechanism based on ASR system. In this research we proposed and implemented a VoWSN based ASR system using Pocketsphinx system. By utilizing base Dic. and LM models of Pocketsphinx, this would result in degrading performance in term of Real-Time limits. This is because it has to consider thousands of words and phrases for each utterance given to it. The solution to this drawback is by customizing Dic. and LM models by stripping them down. Thus, in this paper an algorithm for optimizing Dic. and LM is proposed and implemented. The improvement of VoWSN based ASR using our algorithm is gained by means of:

- Reducing the recognition time due to the elimination of network delay and the reduction in the Dic. size (i.e. reducing searching time).
- Increasing in recognition accuracy (WER is decreased) due to the dedicated words optimized for specific domain.
- Since the optimized Dic. and LM are small in size compared to the default once, reducing Dic. and LM reduces required memory storage.

For more analysis, a comparison study with alternative solution which is VoWSN based streaming system utilizing G.711 and G.729 algorithms is performed to elaborate the advantages of using VoWSN based ASR system in terms of Max. no. of WSN devices that the system can accommodate. Moreover, the two solutions are evaluated when connecting to the Internet. The main conclusions are:

- Results for G.711 and G.729 are approximately equal for voice streaming application in term of maximum number of WSN nodes.
- No. of max. WSN nodes are greatly increased with ASR application compared to voice streaming application with and without Internet. As example with data rate 54M the Max. no. of WSN device is equal to 352 when deploying ASR while it is equal to 19 and 20 with streaming using G.711 and G.729 respectively.

REFERENCES

[1] I. Fathi, Q. I. Ali & J. M. Abdul-Jabbar, "Two Mechanisms to Deploy Real-Time Voice Transmission over Internet of Things (VoIoTs)", International Arab Journal of e-Technology (IAJeT), Vol 5, No.4, 2019.

[2] I. Fathi, Q. I. Ali & J. M. Abdul-Jabbar, "Voice over Wireless Sensor Network (VoWSN) System: A Literature Survey ", International Journal of Information Engineering and Applications Vol.1, No.1, Publication Page: 23-36, Mar. 10, 2018.

[3] S. Phadke, R. Limaye, S. Verma & K. Subramanian, "On design and implementation of an embedded automatic speech recognition system", VLSI Design, Proceedings. 17th International Conference on. IEEE, 2004.

[4] C. Shen, W. Plishker & S. S. Bhattacharyya, "Design and optimization of a distributed, embedded speech recognition system", Parallel and Distributed Processing, IPDPS, IEEE International Symposium on. IEEE, 2008.

[5] F. Sutton, R. Da Forno, R. Lim, M. Zimmerling & L. Thiele, "Demonstration abstract: automatic speech recognition for resource-constrained embedded systems", Proceedings of the13th international symposium on Information processing in sensor networks. IEEE Press, 2014.

[6] R. P. Raghava, RAO & K. M. Lakshmi, "Automatic Speech Recognition for Resource Constrained Embedded Systems", ISSN 2319-8885 Vol.04, Issue. 36, Pages:7701-7708 (2015).

[7] G. Gabor & T. Grosz, "Domain Adaptation of Deep Neural Networks for Automatic Speech Recognition via Wireless Sensors", Journal of Electrical Engineering 67.2 (2016): 124- 130.

[8] M. A. Anusuya & S. K. Katti, "Speech Recognition by Machine: A Review", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 6, No. 3, 2009.

[9] Zh. Tan, B. Lindberg, "Automatic Speech Recognition on Mobile Devices and over Communication Networks", Springer-Verlag London Limited 2008.

[10] A. Sh. Sharma, R. Bhalley, "ASR – A real-time speech recognition on portable devices", 2nd International Conference on Advances in Computing, Communication & Automation (ICACCA), 2016 IEEE.

[11] S. Preethi & B. A. Selvam. "Automatic Speech Recognition System for Real Time Applications." International Journal of Engineering Innovations and Research 2.2 (2013): 157.

[12] A. Rajesh Kumar, and M. Dave. "Implementing a speech recognition system interface for indian languages." Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged languages. 2008.

[13] M. R. Gamit, K. Dhameliya & N. S. Bhatt," Classification Techniques for Speech Recognition: A Review", International Journal of Emerging Technology and Advanced Engineering, Volume 5, Issue 2, February 2015.

[14] P. Sinha, "Speech Processing in Embedded Systems", Springer Science and Business Media, 2010.

[15] Sh. Naziya S. & R. R. Deshmukh, "Speech Recognition System – A Review, IOSR Journal of Computer Engineering (IOSR-JCE), Volume 18, Issue 4, Jul.-Aug. 2016, PP 01-09

[16] H. Tabani," Low-Power Architectures for Automatic Speech Recognition", Doctor of Philosophy, Universitat Polit`ecnica de Catalunya, Spain, 2018.

[17] D. D. Rosário, "Recent advances and challenges in wireless multimedia sensor networks." Mobile Multimedia-User and Technology Perspectives. IntechOpen, 2012.

[18] A. M. E L-Sawy, H. Mahmoud, "A Survey of Wireless Sensor Networks (WSN) Simulators With Multimedia Support", Minia Journal of Engineering and Technology(MJET), Vol.34 , No.2 , 2015.

[19] B. Chen," Audio Recognition with Distributed Wireless Sensor Networks", Master of Science, University of Victoria, 2010.

[20] J. Szurley, A. Bertrand, M. Moonen, & P. Ruckebusch, "Utility based cross-layer collaboration for speech enhancement in wireless acoustic sensor networks", In Signal Processing Conference, 19th European (pp. 235-239). IEEE, 2011.

[21] M. Santos-Pérez, E. González-Parada & J. Manuel Cano-García, "Topic-Dependent Language Model Switching for Embedded Automatic Speech Recognition", Part of the "Advances in Intelligent and Soft Computing" book series (AINSC, Vol. 153.

[22] https://cmusphinx.github.io/wiki.

[23] T. Nilsson, "Speech Recognition Software and Vidispine", Master's Thesis in Computing Science, Umeå University, SWEDEN, April 2, 2013.

[24] D. Z. Ovskiy, "Survey of the Speech Recognition Techniques for Mobile Devices", SPECOM'2006, St. Petersburg, 25-29 June 2006.

[25] M. González1, J. Moreno1 & J. Luis Martínez "An Illustrated Methodology for Evaluating ASR Systems", Springer-Verlag pp. 33–42, 2013.

[26] https://cmusphinx.github.io.

[27] http://www.speech.cs.cmu.edu/tools/lmtool-new.html.

[28] I. Fathi, Q. I. Ali and J. M. Abdul-Jabbar, "Design and Implementation of Real-Time Voice Streaming Evaluation Platform Over Wireless Sensor Network (VoWSN)," International Conference on Advanced Science and Engineering (ICOASE), Duhok, 2018, pp. 233-238.

[29] B. Deshmukh & Sh. Chhatre, "A Technological Survey Of Speech Recognition Techniques", IJAERD-International Journal of Advance Engineering &Research Development, ISNCESR 2017.

[30] https://www.voiptroubleshooter.com/open_speech/american.html.

[31] https://github.com/cmusphinx/pocketsphinx/blob/master/doc/pocketsphinx_continuous.1

[32] Q. I. Ali, "An Efficient Simulation Methodology of Networked Industrial Devices", IEEE SSD08 Conference, Jordan, 2008.

[33] http://www.cs.trincoll.edu/.

[34] V. Smotlacha, "Time Issues in One-way Delay Measurement", Phd thesis, Czech Technical University in Prague, August 2005.

[35] A. M. Alsahlany , " Performance Analysis of VOIP Traffic over Integrating Wireless LAN and WAN Using Different CODECS", International Journal of Wireless & Mobile Networks, Vol. 6, No. 3, June 2014.

[36] Ribadeneira, Alexander F., "An Analysis of the MOS under Conditions of Delay, Jitter and Packet Loss and an Analysis of the Impact of Introducing Piggybacking and Reed Solomon FEC for VOIP." Thesis, Georgia State University, 2007.

الإرسال الصوتي في الوقت الفعلي عبر شبكة المتحسس اللاسلكية (VoWSN) بالاعتماد على تقنية التعرف التلقائي على الكلام (ASR)

إنعام فتحي خضر* **قتيبة ابراهيم علي **** **جاسم محمد عبد الجبار*****

جامعة الموصل \كلية الهندسة\ قسم هندسة الحاسوب

inamfth@gmail.com qut1974@gmail.com drjssm@gmail.com

الملخص

تمثل عملية التعرف على الكلام باستخدام الانظمة المضمنة ذو الموارد المحدودة تحديًا من حيث إمكانية المعالجة وذاكرة التخزين وعرض الحزمة (أو معدل البيانات). لذلك ، في هذا البحث ، تم اقتراح وتنفيذ وتقييم نظام فعال لنقل الصوت في الزمن الحقيق عبر شبكة المتحسس اللاسلكية (*VoWSN*) تعتمد على نظام التعرف التلقائي على الكلام (*ASR*) لاستخدامه في سيناريوهات الطوارئ. يعتمد مبدأ العمل في النظام المقترح على بروتوكول تحويل الفئة (*CTP*) الذي يحول فئة النظام تدريجيًا من نظام *ASR* المعتمد على الشبكة مع نموذج كامل للقاموس واللغة (أي مفردات كبيرة) إلى نظام *ASR* مضمن بالكامل مع قاموس و نموذج لغة مخصصين (أي المفردات الصغيرة). علاوة على ذلك ، تم إجراء دراسة مقارنة بين نظام ال *VoWSN* القائم على نظام ال *ASR* ونظام ال *VoWSN* القائم على نظام البث. تم إجراء هذه المقارنة لتوضيح المكاسب المتحققة عند إرسال نص الإشارة الصوتية بدلاً من إرسال الإشارة الصوتية.
بالإضافة إلى ذلك ، تم تقييم نظام نقل الصوت عبر انترنيت الاشياء (*Voice over IoTs (VoIoTs*) باستخدام الدفق الصوتي أو نظام *ASR* لتقييم أداء النظام عند الاتصال بالإنترنت. تم تنيذ عملية تقييم المقارنة عمليا وباستخدام المحاكاة.

الكلمات الداله :

نقل الصوت عبر شبكة المتحسس اللاسلكي, نظام التعرف التلقائي على الكلام, نظام الدفق, نقل الصوت عبر انترنيت الاشياء , الانظمة المضمنة