# A Ubiquitous Application for Arabic Speech Recognition

Bilal Dendani
*Computer Science Department*
Annaba, Algeria
bilal.dendani@univ-annaba.org

Halima Bahi
*Computer Science Department*
Annaba, Algeria
halima.bahi@univ-annaba.org

Toufik Sari
*Computer Science Department*
Annaba, Algeria
toufik.sari@univ-annaba.org

***Abstract*— The technology of Automatic Speech Recognition (ASR) can be embedded in hand-held devices or reached remotely through the use of server-based speech recognition approaches. We design and implement a ubiquitous speech recognition application for the Arabic language. We select a network-based model for the design and the deployment after a comparison made between speech recognition techniques. We analyze the performance of Arabic speakers through the Arabic speech corpus of isolated words. The experimental results are encouraging to improve the accuracy and efficiency of the proposed system despite the environmental challenges and degradation due to speech coding, the channel itself and the noises affecting speech recognition performance.**

***Keywords—ubiquitous computing; automatic speech recognition, network speech recognition***

## I. Introduction

The prevalence of mobile technologies and devices such as smartphones and wearable devices increases the interest towards speech recognition applications in everyday lives. The technology behind that tremendous progress is the ubiquitous computing which encourages automatic speech recognition to be accessible everywhere, transparently without thinking and every day. Therefore, challenges will affect the performance of classical automatic speech recognition (ASR) systems. The main motivation of this work is the investigation of the remote ASR systems and their use as ubiquitous systems.

An automatic speech recognition system converts speech captured signal to a sequence of words. The ASR systems are based on statistical pattern recognition models particularly Hidden Markov Models (HMMs) [1]. From a speech signal, we extract feature vectors that represent the observation data Y. The word sequence Ŵ is obtained through Bayesian decision rule:

$$\hat{W} = argmax\,(P(W|Y) = argmax\ P(W) * P(Y|W) \qquad (1)$$

The P(W) is the prior probability of observing some specified word sequence W and is given by a language model.

P(Y|W) is the likelihoods, probability of observing the speech data Y given the word sequence W, it is determined by an HMM acoustic model.

Several research works focused on Arabic continuous speech recognition. In [2], an Arabic built speech recognition system based on Sphinx 4 open source. The resulted system used acoustic model, language model and the phonetic dictionary for the Arabic digits (0...9). Abushariah et al. [3] designed and implemented a high-performance natural speaker independent ASR system. The Main parts of ASR system are influenced by the ubiquitous computing technology.

Mark Weiser introduced the concept of ubiquitous computing in (1991) [4] where services and information are accessible by users everywhere, at any time. The ubiquitous technology reaches several fields and applications, such as healthcare, smart houses and language learning. In [5], a ubiquitous health monitoring solution and well-being for the overwhelming problems of the healthcare system is presented. The proposed system integrated wearable and environment sensors to survey weight, physical and heart activities. Kumar [6] developed a system which controls and monitors smart house environment. The system contains an android application which communicates with a micro web server using voice activation. Another example introduced in [7], a cloud-assisted speech recognition service (CSR) based on distributed speech recognition (DSR) model. The CSR framework uses servers in the cloud to improve the accuracy and the speed of ASR. Moreover, a ubiquitous ASR application for Mixtec Language used for translation between Mixtec and Spanish as in [8].

Arabic ubiquitous speech recognition systems are almost nonexistent. Qidwai and Shakir [9] designed a ubiquitous, effective Arabic voice control device for disabled people based on continuous speech recognition. The central motivation for this work came from the rarity of researches related to ubiquitous Arabic speech recognition, encouraging us to work on the design and implementation of such system.

We propose a ubiquitous speech recognition system based on network speech recognition (NSR) model for the Arabic language. The system provides a client/server design architecture. The client-side mobile application captures audio data and transfers it to a connected server which process data across wireless communication. This application will be used for developing a future large vocabulary ubiquitous speech recognition system for the Arabic language.

The remainder of this paper is organized as follows: Section II describes the different speech recognition models used for the deployment of ASR applications, their advantages and disadvantages. In section III, a ubiquitous computing application based on (NSR) is described. In section IV, practical issues are presented. Finally, a conclusion is drawn.

## II. Automatic Speech Recognition Models

### A. ASR Models

There are three approaches to deploy ASR applications: Network Speech Recognition (NSR), Distributed Speech Recognition (DSR) and Embedded Speech Recognition (ESR) models. For embedded speech recognition (ESR), the demand of resources is considered as an obstacle because of the scarred consumer devices resources, this downside represents exactly the benefit of a remote automatic speech recognition that has an advantage in term of available resources.

The speech signal is always captured on the client side and the application can reside on the client or on the server, however, the distribution of other remaining components follows the three approaches according to some factors including the resource availability, the complexity of the components and the application location. The deployment of an ASR in devices and networks enables a flexible architecture.

Fig.1 represents the typical architecture of an ASR system, from speech signal capturing to feature vectors extraction front-end to back-end speech recognition decoding.

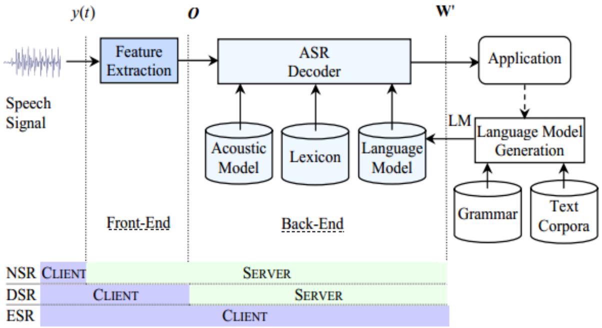

Fig. 1. Typical ASR architecture (from [10].)

The front-end processing is less resource demanding, however, the back-end requires more resources because of a huge number of parameters calculated for the acoustic model. The HMM-based back-end needs more computational resources like memory and CPU speed. The decoding process also needs a substantial amount of CPU resources and a speed memory access for searching the most likely sequence of words.

### B. Speech Recognition Application Deployment

The deployment of ASR application in an embedded system defer to deployment in a network-based architecture. The obstacle of the lack of resources in embedded speech recognition (ESR) represents the benefit of using remote ASR which can be a distributed or a network according to the position of feature extraction. Speech signal quality and (noise and channel) robustness are important parameters for choosing DSR while the wide deployment of high-quality speech coders makes NSR a favourite.

In our application, we choose the architecture of network speech recognition.

## III. Ubiquitous Computing Application based on NSR Model

One of the motivations for using the architecture of NSR instead of DSR and ESR is the simplicity to update the ASR systems at the server side. In [3], authors conclude that NSR performance is comparable to DSR when using the AMR modes coder for speech compression. NSR enables plug and play of the ASR system on the server side without changes on the client devices.

Network speech recognition (NSR) is identified by the location of both feature extraction and ASR search in the server-side while the speech signal is captured in the client-side as shown in Fig .2. Sometimes NSR is referred to a cloud speech recognition system. Network and cloud-based speech recognition systems can be used in developing regions where the terminals are low-resources cellular phones according to [11]. No need for increased resources in client-side because the central server is responsible for feature extraction and searching process. Fig. 2 presents a ubiquitous network speech recognition system for the Arabic language inspired from [8].

One of the disadvantages caused by this approach is the performance degradation of the recognition process due to the use of low bit rate codecs for encoding speech, which becomes more severe when data transmission errors occur and in case of noise background [12]. Another issue of this mode is that it should serve requests that come simultaneously from clients. To overcome these limitations, we consider the G.711 codec as the speech compression algorithm.

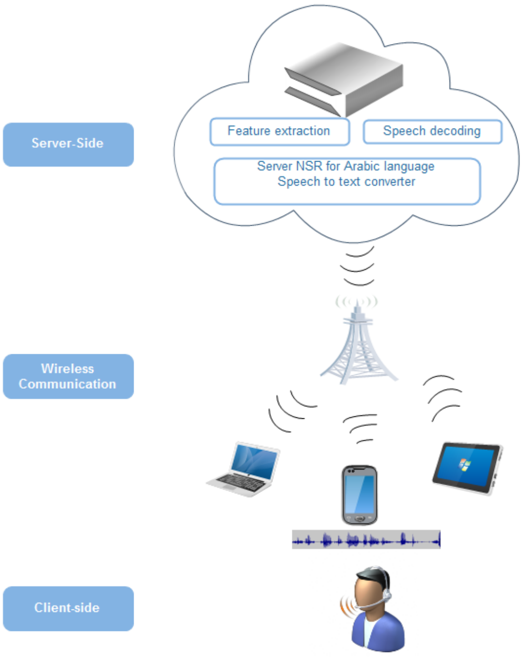

Fig. 2. A ubiquitous network-based ASR system for Arabic

This application contains:

- The core ASR system deployed in the server side using the open source Sphinx 4 and models including acoustic, language model and the dictionary.
- Client mobile application that captures the user's voice and sends it to the server to be processed.

### A. Client Side

The client-based application is implemented using Android Studio 2.3. In Fig. 3, the user connects to the server-side then records his / her voice. The connection is made by java sockets.

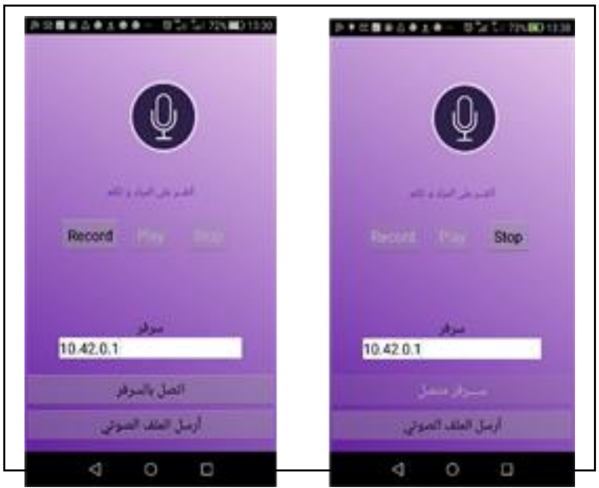

Fig. 3. A ubiquitous network-based ASR system for Arabic

This mobile application mainly

- Connects to the server which recognizes the audio data,
- Captures the speech of the client and compresses it,
- Sends speech signal across wireless communication.

### B. Server Side

The server contains the speech engine that recognizes the transmitted voice from the mobile client according to speech models (acoustic model, language model and a phonetic dictionary). This is done using Sphinx 4 open source [13].

The server used for building our ubiquitous Arabic speech recognition system is a laptop HP dv6, processor Intel CORE I 5 and a memory size 8 GO. The server recognizes speech and returns the result to the client mobile app.

## IV. Practical Issues

The system was implemented as previously described. We evaluate the performances of this system in real life conditions.

### A. Speech Dataset

The used corpus is an Arabic speech corpus for isolated words. The corpus has been developed by the Department of Management Information Systems, King Faisal University [14]. It contains 9992 recorded utterances of 50 speakers pronouncing 20 words.

### B. Speech Recognition

The speech recognizer was developed using the CMU Sphinx toolkit [13]. To build an automatic speech recognizer two components are required: Acoustic models and language model. In this research, acoustic models are built on the basis of the Mel frequency cepstral coefficients (MFCC) representation of phonemes.

### C. Performances

To test the ASR system performances, we compute the Accuracy.

$$Accuracy = 1 - WER \quad (2)$$

The word error rate (WER) score is computed by comparing a reference transcription with the transcription outputted by the speech recognizer. From this comparison, it is possible to compute the number of errors which typically belong to three categories: insertions (*I*), deletions (*D*) and substitutions (*S*).

$$WER = \frac{I + D + S}{N} \quad (3)$$

*N* is the number of words in the reference transcription. The performance of the proposed network speech recognition application is depicted in Fig 4. Compared to the offline test, there is a slightly a neglected deterioration in speech recognition performance. The accuracy of speaker11 decreases from 90% to 88.33% and that of the speaker2 decreases from 60% to 58%. Most of the other speakers preserve their performance. The better performance is for speaker11 (88.33%), speaker7 (76.66%) and speaker13 (76.66%) explained by the effect of the intelligible quality of corresponding samples. The performance degradation is due to the unintelligible samples and to the network wireless impairments caused by the packet loss, wireless impediments and the jitter factor.

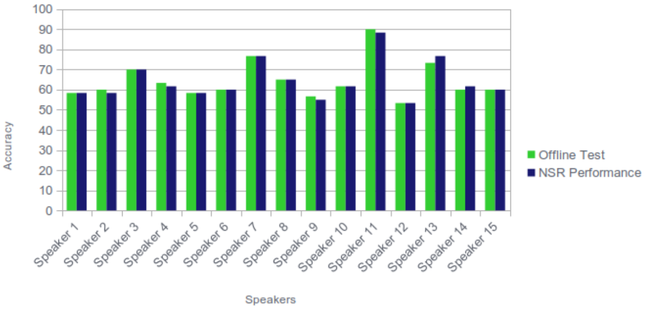

Fig. 4. Performance of the ASR system

The accuracy over all the speakers reaches 64.33%.

# V. Conclusion

A ubiquitous speech recognition systems for the Arabic language is critical for billions of internet users. The deployment of ASR application in embedded systems defers to deployment in a network-based architecture. The obstacle of the lack of resources in an embedded speech recognition model represents the benefit of using remote ASR which can be distributed or network according to the position of features extraction. Speech signal quality and (noise and channel) robustness are important parameters for choosing DSR while the wide deployment of high-quality speech coders makes NSR a favourite.

This work presents the design and implementation of such a system for mobile devices. The proposed system is based on a network speech recognition model.

As a starting point, this paper presents some results that will encourage as to

- Improve the performance, accuracy and efficiency of the ubiquitous ASR system.
- Propose a future distributed speech recognition (DSR) for the Arabic language which enables speech recognition application via the web and compares it with network speech recognition (NSR) application in terms of efficiency and accuracy.
- Build a ubiquitous Arabic speech recognition system which recognizes words according to a large vocabulary.
- Allow simultaneous and concurrent clients connections.
- Allow students to track and follow lessons made by teachers from a speech mobile-based interface.

# *References*

[1] H. Bahi and M. Sellami, “Combination of vector quantization and hidden Markov models for Arabic speech recognition,” in Proceedings ACS/IEEE International Conference on Computer Systems and Applications, pp. 96-100, 2001.

[2] H. Satori, M. Harti, and N. Chenfour, “Introduction to Arabic Speech Recognition Using CMUSphinx System,” Apr. 2007.

[3] M. Abushariah, R. N. Ainon, R. Zainuddin, M. Elshafei, and O. Khalifa, Natural speaker independent arabic speech recognition system based on HMM using sphinx tools. 2010.

[4] M. Weiser, The Computer for the 21st Century, vol. 265. 1991.

[5] M. Milošević, M. T. Shrove, and E. Jovanov, “Applications of Smartphones for Ubiquitous Health Monitoring and Wellbeing Management.” 2011.

[6] S. Kumar, “Ubiquitous Smart Home System Using Android Application,” Int. J. Comput. Networks Commun., vol. 6, no.1, 2014.

[7] Y.-S. Chang, S.-H. Hung, N. J. C. Wang, and B.-S. Lin, “CSR: A Cloud Assisted Speech Recognition Service for Personal Mobile Device,” in 2011 International Conference on Parallel Processing, 2011, pp. 305-314.

[8] S. O. Caballero-Morales and F. Trujillo-Romero, “Automatic speech recognition of the Mixtec language: An ubiquitous computing application,” in Proc., 23rd International Conference on Electronics, Communications and Computing, pp. 98–103,2013.

[9] U. Qidwai and M. Shakir, “Ubiquitous Arabic voice control device to assist people with disabilities,” in 2012 4th International Conference on Intelligent and Advanced Systems (ICIAS2012), 2012, pp. 333–338.

[10] Z.-H. Tan and I. Varga, Network, Distributed and Embedded Speech Recognition: An Overview, in Automatic Speech Recognition on Mobile Devices and over Communication Networks, London, 2008, pp. 1–23.

[11] A. Kumar, A. Tewari, S. Horrigan, M. Kam, F. Metze, and J. Canny, “Rethinking Speech Recognition on Mobile Devices,” Proc. Int. Workshop Intell. User Interfaces Dev. Reg. IUI4DR, Feb. 2011.

[12] A. Schmitt, D. Zaykovskiy, and W. Minker, “Speech recognition for mobile devices,” Int. J. Speech Technol., vol. 11, no. 2, pp. 63–72, Jun.2008.

[13] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf and J. Woelfel, Sphinx-4: a flexible open source framework for speech recognition, Technical Report, Sun MicroSystems Inc., 2004.

[14] A. Alalshekmubarak and L. S. Smith, “On improving the classification capability of reservoir computing for arabic speech recognition,” *LNCS*, vol. 8681, pp. 225–232, 2014.