A Smartphone-Based Multi-Functional Speech-To-Text Transcription System

Abayomi O. Agbeyangi[*] & Adam B. Olorunlomerue[+]
[*]Department of Computer Science, Chrisland University, Abeokuta, Ogun State, Nigeria
[+]Department of Computer Science, Federal Polytechnic, Ede, Osun State, Nigeria
aagbeyangi@chrislanduniversity.edu.ng, olorunlomerue.adam@federalpolyede.edu.ng

ABSTRACT

The design and implementation of speech-to-text and text-to-speech systems to cater for the need of disabled and language inefficient had been seen to be highly encouraging in recent time. Its use with different speech recognition technologies is with some challenges. In this paper, a smartphone-based speech-to-text system with LCD display is presented. In the design, a voice recognition module builds into the system perform the speech recognition task while a dedicated Android application does the speech transcription. The recorded voice is transcribed to English sounds through the android application and matched to the stored words in the database. The converted text is then sent wirelessly from the Android application to the LCD display. The results from various testing show that the system performs excellently. Although there are challenges noted for further study.

Keywords: Speech-to-text, speech recognition, speech synthesis, speech transcription, LCD Display

iSTEAMS Proceedings Reference Format
Abayomi O. Agbeyangi & Adam B. Olorunlomerue (2019): A Smartphone-Based Multi-Functional Speech-To-Text Transcription System. Proceedings of the 15th iSTEAMS Research Nexus Conference, Chrisland University, Abeokuta, Nigeria, 16th – 18th April, 2019. Pp 165-174
www.isteams.net - DOI Affix - https://doi.org/ 10.22624/AIMS/iSTEAMS-2019/V15N1P17

1. INTRODUCTION

The use of speech recognition technology that is able to capture spoken words by human using microphone seems to span several decades. The spoken words recorded are subjected to the speech recognizer, which matched the words against dictionary words in the database and output the recognized words. Basically, the effectiveness of the technology depends on some factors such as users, vocabularies, and the environment. This technology started around the 1940s, the first was in 1952 called Audrey at the bell labs to recognize a digit in a noise-free environment (Shadiev et al., 2014; Joshi et al., 2017).

According to Bansal et al. (2018), speech-to-text translation which can be seen as a pipeline of automatic speech recognition and machine translation has many potential applications for low-resource languages. Some noted scenario was for language documentation, for unwritten or endangered source language, in crisis relief where aid workers need to respond to request or calls in a different language (foreign language). In these scenario raised by Bansal et al, the use of speech-to-text system will adequately address the situation. Many of the several applications of automatic speech recognition system are to transcribe speech such as talks, lecture, presentations, broadcast news and phone conversation. The speech-to-text engine can also provide data entry options for blind, deaf, and other physically handicapped users.

Furui et al (2004) argued that "although speech is the most natural and effective method of communication between human beings, it is not easy to quickly review, retrieve, and reuse different forms of speeches if they are simply recorded as audio signal". Thus, transcribing speech mainly to text is expected to become a crucial capability for the future information technology era. Most speech recognition systems require voice enrolment or training where an individual speaker speaks into the system. The system analyzes the specific voice and uses that to fine-tune the recognition of the speech in order to increase accuracy. Thus, automatic speech recognition systems (ASR) operate primarily in two stages (see Figure 1 for illustration); training phase where the learning of the reference patterns of the different speech sounds (e.g. phonemes, phrases, words) takes place.

Pattern recognition phase is the second phase where an unknown input pattern from the training is identified by using the set of references (Khilari & Bhope, 2015). Other systems that did not support voice training are referered to as speaker independent systems (Nguyen et al., 2010). Basically, voice recognition encompasses the task of identifying the speaker, other than what they are saying, this simplifies the task of translating speech in systems that have been trained on a specific individual's voice. The field in recent time benefited from advances in deep learning and big data. This has brought a surge in many worldwide industries adopting a variety of deep learning methods in designing and deploying speech recognition systems.

It has been confirmed that the evaluation of the performance of speech recognition systems is by accuracy and speed. Its accuracy is rated by word error rate (WER) and Command Success Rate (CSR) while speed is on the real-time factor (Gerbino et al., 1993; Pieraccini, 2012). Speech recognition by machine is a very complex task due to individual accent, words articulation, pronunciation, pitch, roughness, nasality, volume and speed. Background noise, echoes and electrical characteristics can also distort speech recognition.

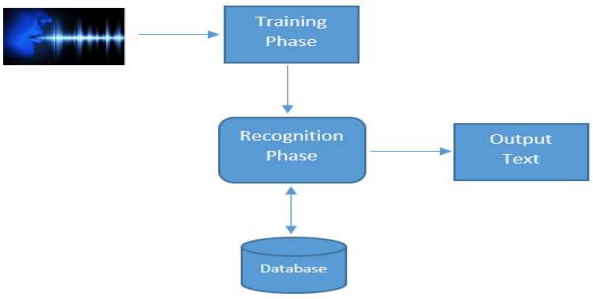

Fig. 1: Automatic Speech Recognition Phases

This study seeks to develop a speech-to-text system that is capable of listening to human's voice through a mobile phone microphone, convert the voice to text using an Arduino application and a hardware display system via a Bluetooth communication protocol and then displayed the spoken speech as a text on an LCD display. The first stage is the voice/speech acquisition, where the microphone takes the speech heard and send it to the android application for transcription. The second stage is the transcription where the recorded word is transcribed to English sounds through the android application with the support of google voice engine and then matched to the stored words in the database. The next state is the communication stage where the converted text is then sent wirelessly to the Arduino-based hardware system via Bluetooth. The last stage is where the converted text is then displayed on the LCD display.

The remaining part of the paper is organized as follows: section 2 discusses related works; section 3 examines material and methods used; section 4 discusses the results, while section 5 concludes the paper.

2. RELATED WORKS

In literature, many works has been reported on speech-to-text recognition (Reddy & Mahender, 2013; Zhu et al., 2014; Upadhyay & Chavda, 2014; Chern et al., 2017) with several technologies ranging from the traditional acoustic and hidden Markov models (Juang & Rabiner, 1991) to today's modern machine learning (Deng & Li, 2013). In hidden Markov model, the output is a hidden probabilistic function of a state which cannot be deterministically specified. Thus, the speech recognition system using Markov model generally assume that the speech signal is as a result of some message encoded as a sequence of one or more symbols.

In Upadhyay & Chavda (2014), the design of a GUI application with Raspberry Pi was reported. The device goal is to enable blind people to search for anything. The system recognizes the speech spoken and represent optimized data that adequate transcribed the meaning of the speech to text. A similar work was reported in Niculescu et al. (2014) where a multimodal dialogue system to assist touristic visiting Singapore was developed. It is an Android mobile application called SARA that provides information about local attractions, sight-seeing, direction, restaurants, and transportation services. To communicate with the system, the user either speaks, text or scanned the QR code. Its output also can either be speech or text.

Reddy & Reddy (2013) presented a system to aid the blind navigation and deaf hearing ability using speech recognition. The system help detects and avoids obstacle for the blind with the ability to communicate using speech recognition module and LCD display. Ultrasonic sensor and buzzer were used to detect and avoid an obstacle while Flex sensor interfaced to a speaker was used to aid the dump to express their feelings. Pressure sensor was also incorporated in the system to help the disabled and visually impaired in case they fall. The sensor activates the GSM Module in the system and sends SMS to their guardian. The system is robust as it has lots of functional tools to aid the physically challenged.

Ranchal et al. (2013) in their work on speech recognition for captioning and lecture transcription within the classroom with the basic aim of assisting the students to convert oral lectures to text automatically. In the study, two different approaches to SR-mediated lecture acquisition (SRmLA) was considered and employed using conventional educational technology in a contemporary university lecture room. Captioning (RTC) of the instructor's lecture speech was done using a client-server application to facilitate instant lecture viewing during class via a projection screen and also directly to the students' laptops. Post-lecture transcription (PLT) was also employed in the study to digitally record the instructor's lecture to provide automatic transcription for students to view online or download after class. Experimental results from the study showed that a greater word recognition accuracy and flexible use of multimedia class notes were recorded. As seen from the results in the research, the system will be very useful for students with special needs and non-native English language to obtain class notes. A similar work to this was also reported in Chern et al. (2017), a smartphone-based (SmartHear) hearing assistive speech recognition system users to enhanced listening clarity in the classroom.

Another study worthy of note was reported in Muthuselvi & Saravanan (2014), where an embedded system for home automation to detect and recognize human voice commands through speech recognition was implemented. The voice command was to toggle respective loads connected to the system. Voice input detected will be analysed by the speech recognition module in a way similar to what was reported in Reddy & Reddy (2013). The resulting outputs are then displayed on an LCD to signify the system state.

In Hannun et al. (2014), an end-to-end deep learning model for speech recognition task was reported. The approach was significantly simpler if compare with the traditional speech recognition systems (acoustic models and Hidden Markov Models). It was an improvement on the old method as there is no need for hand-designed components to model reverberation and background noise but instead, the system learns a robust function to such effects. Thus, no need for phoneme dictionary as required for hidden Markov models. In the approach, a well-optimized RNN training system with multiple GPUs was employed. Experimental results from the system (Deep Speech) outperforms other similar studies.

Other works similar to this study are: the work of Cox et al. (2002), an experimental speech recognizer system (TESSA) to aid transactions between deaf and Post Office clerk that transcribe the speech to sign language; Reddy & Mahender (2013), an online speech-to-text system which acquires speech through a microphone and processes speech in order to recognize spoken text. The work is relatively similar to our work as it directly acquires and converts the speech to text. But differ from ours in the method of output communication and display. Our system has an additional hardware part which accepts the transcribed text wirelessly from the android app and displayed on an LCD display.

3. MATERIAL AND METHODS

It has been established that a spontaneous speech ability that is able to handle different words and natural language speech being run concurrently must be the goal of a good speech recognition system. Since different speakers have their special voices owing to unique nature and personality. In the development of our system, we realised that a speaker model to adopt will be independent of the speaker since we want the system to be used by different peoples. Thus, we adopt the speaker independent model approach.

In summary, the highlighted phases the system pass through to display the spoken words is given below:

(i) Voice/speech Acquisition: The microphone takes the speech heard and sends it to the android app for transcription;
(ii) Transcription: The recorded words are transcribed to English sounds through the android application with the support of Google voice engine which matched to the stored words its database;
(iii) Communication: The converted text is then sent wirelessly to the Arduino-based hardware system via the Bluetooth;
(iv) Display: The converted text is then displayed on a 16x2 LCD module.

This is further illustrated in Figure 2.

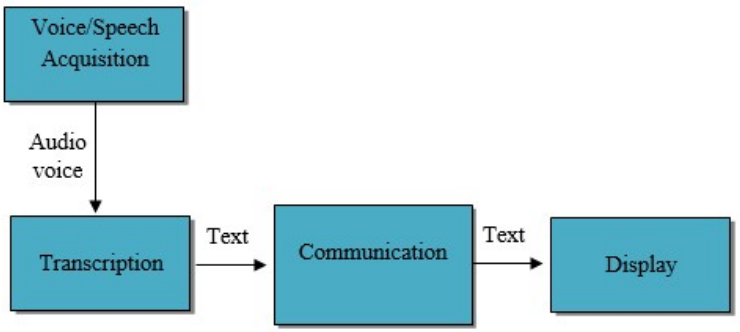

Fig. 2: The Speech-to-Text Development Modules

For the voice recognition task in the transcription module, we use acoustic model (Lamel et al., 2002; Wang et al., 2003). The model is used in automatic speech recognition (ASR) to represent the relationship between the audio signal and the phonemes in the speech. The model works by taking audio recordings of speech with their text transcriptions using software to create statistical and appropriate representations of the sounds that make up each word in the speech. We decided to use the model due to the challenges of training involved in other models. The overall architecture of the system is shown in Figure 3.

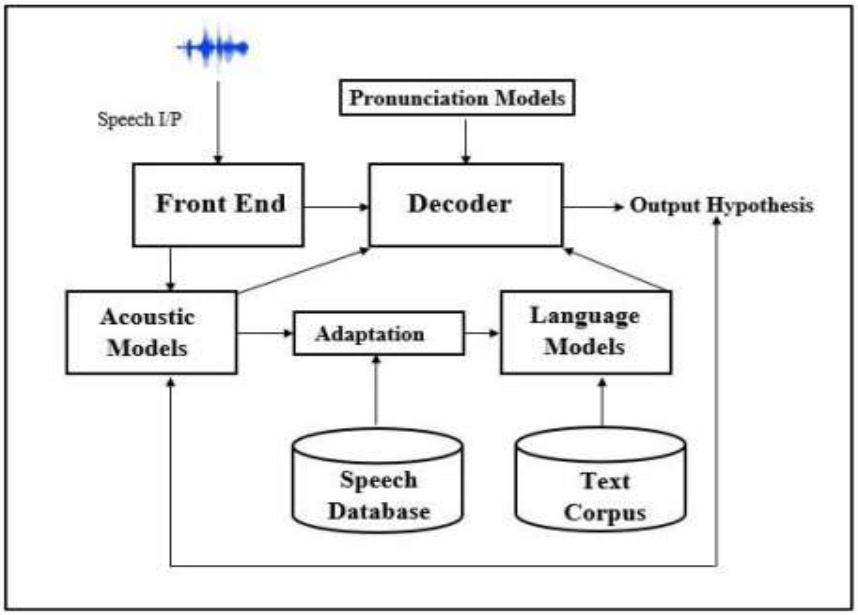

Fig. 3: The System Architecture

The hardware part of the system handles the output text display. In the design, the circuit involves the alignment of Arduino, LCD and Bluetooth module together as illustrated in Figure 4. The Arduino board comes with an inbuilt female pin header which has connections to its input/output pins, TX (transmit signal), RX (receive signal), ground pins and voltage pins, a male pin header can be mounted on motor drivers that comes with the same alignment of pins so it can be overlaid on the Arduino. Thus, there is no need for wires to join between them. In the process, the voice inputs are processed to check for appropriate inputs from the software application. If the inputs are in proper format then a text command is generated which would be sent to the microcontroller to be displayed on the LCD module.

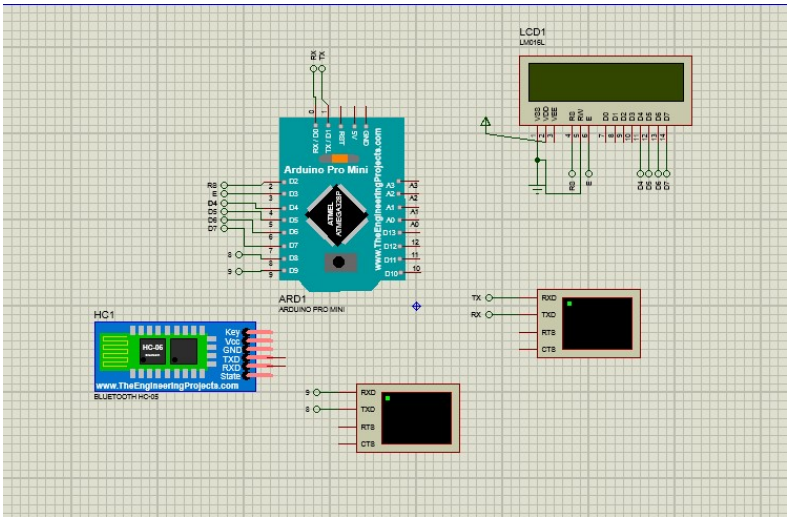

Fig. 4: The Hardware part circuit diagram Simulation

4. RESULTS AND DISCUSSION

The process of testing the functionality of the system was divided into two parts:

(i) The speech recognition; and
(ii) The hardware part

We decided to undergo a qualitative approach to analyse the performance of the system since we assumed doing a quantitative analysis might be difficult. Consequently, it was noted that with speech it is almost impossible to fully control the variables because each person has a different voice and also it might even be difficult to pronounce just a phrase in the exact same manner two times keeping the background noise constant. The background noise will vary constantly.

So we decided to use three persons with different quality of spoken English and the test was conducted in a room with little background noise (see Figure 7). The experimental result obtained shows very little variation in the output produced by the system. Thus, justifying our expectation of the system performance. Figure 5 shows the full packaging of the hardware part while Control Application is shown in Figure 6.

Fig. 5: The Complete Hardware

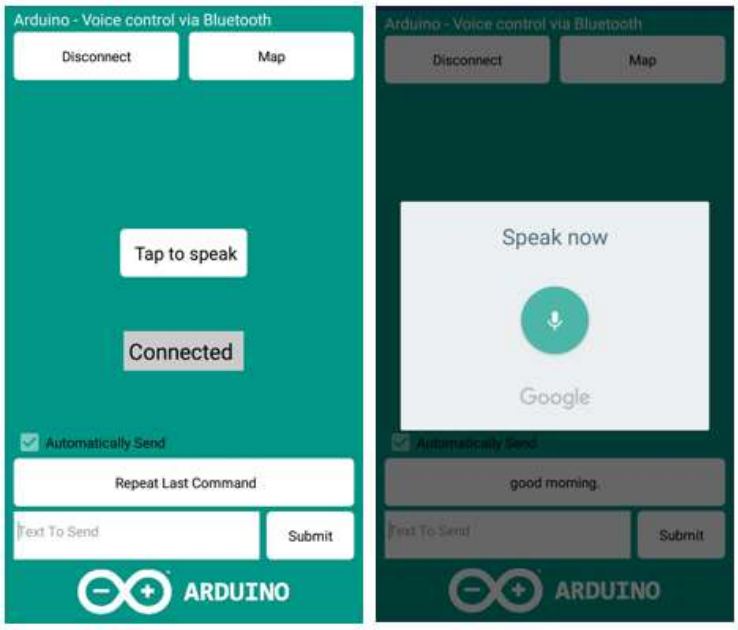

Fig. 6: The Android Control Application Interface

Fig. 7: The System Ouput During Testing

5. CONCLUSION

The study demonstrates the enormous features of speech to text system using acoustic model. The developed system is believed to be low cost and user-friendly. It will assist people facing hearing difficulty and others who might find it useful. Its use can also be extended to lecturers in the classroom to record the transcript of their lecture. For further research, we intend to increase the LCD display size in order to accommodate more text at a time. We also plan to make the system more intelligent by adding a feature that will enable it to guess what the speaker intends to speak, rather than what was actually spoken, as people often make mistake during pronunciation particularly due to dialect.

REFERENCES

(i) What is speech recognition? - Definition from WhatIs.com. (2019). SearchCRM. Retrieved 11 January 2019, from https://searchcrm.tech target.com/definition/speech-recognition

(ii) Bansal, S., Kamper, H., Livescu, K., Lopez, A., & Goldwater, S. (2018). Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. *arXiv preprint arXiv:1809.01431*.

(iii) Furui, S., Kikuchi, T., Shinnaka, Y., & Hori, C. (2004). Speech-to-text and speech-to-speech summarization of spontaneous speech. *IEEE Transactions on Speech and Audio Processing*, *12*(4), 401-408

(iv) Nguyen, P., Tran, D., Huang, X., & Sharma, D. (2010). Automatic classification of speaker characteristics. In *International Conference on Communications and Electronics 2010* (pp. 147-152). IEEE.

(v) Gerbino, E., Baggia, P., Ciaramella, A., & Rullent, C. (1993). Test and evaluation of a spoken dialogue system. In *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 2, pp. 135-138). IEEE.

(vi) Pieraccini, R. (2012). *The voice in the machine: building computers that understand speech*. MIT Press.

(vii) Joshi, S., Kumari, A., Pai, P., Sangaonkar, S., & D'Souza, M. (2017). Voice Recognition System. *Journal for Research*, *2*(11).

(viii) Shadiev, R., Hwang, W. Y., Chen, N. S., & Huang, Y. M. (2014). Review of speech-to-text recognition technology for enhancing learning. *Journal of Educational Technology & Society*, *17*(4), 65-84.

(ix) Reddy, B. R., & Mahender, E. (2013). Speech to text conversion using android platform. *International Journal of Engineering Research and Applications (IJERA)*, *3*(1), 253-258.

(x) Zhu, Z., Branzoi, V., Wolverton, M., Murray, G., Vitovitch, N., Yarnall, L., ... & Kumar, R. (2014). AR-mentor: Augmented reality based mentoring system. In *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (pp. 17-22). IEEE

(xi) Chern, A., Lai, Y. H., Chang, Y. P., Tsao, Y., Chang, R. Y., & Chang, H. W. (2017). A smartphone-based multi-functional hearing assistive system to facilitate speech recognition in the classroom. *IEEE Access*, *5*, 10339-10351

(xii) Juang, B. H., & Rabiner, L. R. (1991). Hidden Markov models for speech recognition. *Technometrics*, *33*(3), 251-272.

(xiii) Deng, L., & Li, X. (2013). Machine learning paradigms for speech recognition: An overview. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(5), 1060-1089

(xiv) Upadhyay, M. S. K., & Chavda, M. V. N. (2014). 'INTELLIGENT SYSTEM BASED ON SPEECH RECOGNITION WITH CAPABILITY OF SELFLEARNING'. *International Journal For Technological Research In Engineering*, *1*(9)

(xv) Niculescu, A. I., Jiang, R., Kim, S., Yeo, K. H., D'Haro, L. F., Niswar, A., & Banchs, R. E. (2014, August). SARA: Singapore's Automated Responsive Assistant, a multimodal dialogue system for touristic information. In *International Conference on Mobile Web and Information Systems* (pp. 153-164). Springer, Cham

(xvi) Reddy, G. Y., & Reddy, M. R. (2013). Design of ultrasonic spectacles, flex sensor based voice generation and speech to text (data) conversion techniques for physically disable people. *Int. J. Rev. Electronics Communication Engineering (IJRECE)*, *1*, 12-22

(xvii) Venayagamoorthy, G. K., Moonasar, V., & Sandrasegaran, K. (1998). Voice recognition using neural networks. In *Proceedings of the 1998 South African Symposium on Communications and Signal Processing-COMSIG'98 (Cat. No. 98EX214)* (pp. 29-32). IEEE

(xviii) Ranchal, R., Taber-Doughty, T., Guo, Y., Bain, K., Martin, H., Robinson, J. P., & Duerstock, B. S. (2013). Using speech recognition for real-time captioning and lecture transcription in the classroom. *IEEE Transactions on Learning Technologies*, *6*(4), 299-311

(xix) Muthuselvi, G., & Saravanan, B. (2014). Real Time Speech Recognition Based Building Automation System. *ARPN Journal of Engineering and Applied Sciences*, *9*(12), 2831-2839.

(xx) Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., ... & Ng, A. Y. (2014). Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

(xxi) Cox, S., Lincoln, M., Tryggvason, J., Nakisa, M., Wells, M., Tutt, M., & Abbott, S. (2002). Tessa, a system to aid communication with deaf people. In *Proceedings of the fifth international ACM conference on Assistive technologies* (pp. 205-212). ACM

(xxii) Khilari, P., & Bhope, V. P. (2015). A review on speech to text conversion methods. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, *4*(7)

(xxiii) Reddy, B. R., & Mahender, E. (2013). Speech to text conversion using android platform. *International Journal of Engineering Research and Applications (IJERA)*, *3*(1), 253-258.

(xxiv) Shadiev, R., Hwang, W. Y., Chen, N. S., & Huang, Y. M. (2014). Review of speech-to-text recognition technology for enhancing learning. *Journal of Educational Technology & Society*, *17*(4), 65-84.

(xxv) Chern, A., Lai, Y. H., Chang, Y. P., Tsao, Y., Chang, R. Y., & Chang, H. W. (2017). A smartphone-based multi-functional hearing assistive system to facilitate speech recognition in the classroom. *IEEE Access*, *5*, 10339-10351

(xxvi) Lamel, L., Gauvain, J. L., & Adda, G. (2002). Lightly supervised and unsupervised acoustic model training. *Computer Speech & Language*, *16*(1), 115-129.

(xxvii) Wang, Z., Schultz, T., & Waibel, A. (2003). Comparison of acoustic model adaptation techniques on non-native speech. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).* (Vol. 1, pp. I-I). IEEE