The Importance of Human Speech Data: Recognizing Human Input Speech in Speech-to-Text Systems

Emmanuel Odonkor Teye-Kofi
emmanuel.odonkor@ashesi.edu.gh

ABSTRACT

The increased availability of Speech-to-Text systems today has led to its utilization in various sectors. The processes and computations of these speech-to-text systems makes them act as if they work independently without the provision of human speech as input data in the first place. We provide evidence from Speech-to-Text systems in the field of health and education that the case of Speech-to-Text systems working independently is invalid because they fully depend on human speech as input data before they can function. We also elaborate on the idea that Speech-to-Text systems used in any sector cannot deliver its functions if human speech as input is not provided. The evolution of natural language processing is also discussed to highlight on ELIZA, the first natural language processing system and how over the generations, improvement have been made to it through the invention of advanced Natural language processing systems. Contributions of NLP in other areas are discussed to unveil how Natural language processing as a subfield of Artificial intelligence is helping to make the world a better place.

KEYWORDS

Natural Language Communication, Natural language Processing(NLP), Humanoid, Intelligence, Symbolic reasoning, Virtual personal assistants, Artificial Intelligence, SCRIPT

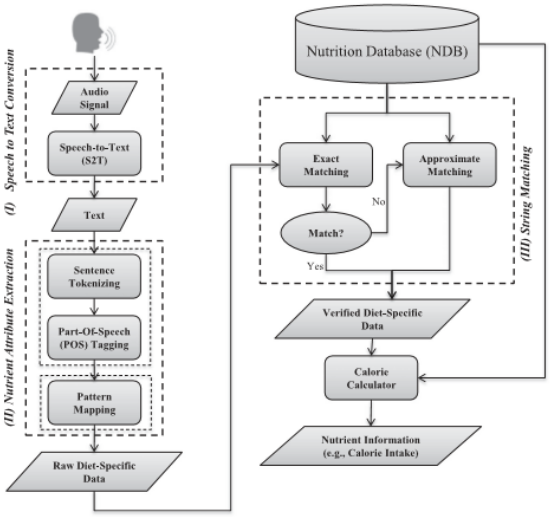
CS313 Reference Format:
Emmanuel Odonkor Teye-Kofi. 2020. The Importance of Human Speech Data: Recognizing Human Input Speech in Speech-to-Text Systems. In *Proceedings of (CS313).* , 6 pages.

1 INTRODUCTION

In this dispensation, there is vast utilization of Natural language processing systems in existence since the emergence of ELIZA in 1966 [6]. More specifically, Speech-to-Text systems have been known to assist humans in the sector of health and education, and have proven to be efficient in terms of performance when delivering their roles [3, 10]. Technically speaking, people are amazed at how these systems can receive human voice input, process them to clearly understand what they are supposed to do, and eventually deliver the right output. Due to the great support, and too much attention attributed to how Speech-to-Text systems marvelously work, there are perceived to work independently and operate without the intervention and recognition of human speech as input data when delivering their functions. Therefore as a result of this, the point of understanding that it was the contribution of human speech input in the first place to give these systems a head start to deliver their roles is forgotten.

Speech-to-Text systems are NLP systems that are embedded with voice recognition technology to enable them to receive human speech, process these input, and convert them to text for further processing by computers [10]. Looking at current systems, the mobile-based Speech-to-Text system, Speech2Health, uses human speech data and based on that, effectively monitors people's nutrition by accurately computing their calorie intake values [10]. Another prior research highlights the Nutritional Dialogue System, another Speech-to-Text system that extracts food concepts from people via speech as input, and based on it, determines the nutrition facts and assesses its intake in people [7]. These Speech-to-Text systems have proven to be helpful in health sectors by helping individuals keep track of their diet intake. Most importantly, due to its intervention to speed up nutritional monitoring in humans, they have been recognized to enhance scalability and clinical utility of mobile nutrition monitoring but no credit was given to human speech that helped to enable these devices to operate. The above reputation given to these systems would not have been possible without the provision of human speech as input. These systems cannot detect nutrition facts if humans have not given in their nutritional data in the first place. The Speech2Health System would be addressed in detail to highlight the relevance of human speech as input data.

1.0.1 *Fig1:Role of Human Speech in Speech2Health System.*



CS313, 2020, Ashesi University, Ghana

Fig 1 shows the schematic implementation of the Speech2Health System. The idea behind this diagram is to stimulate the role and importance of the human speech to enable the system perform its functions. Without the human voice given as input, this systems would not be able to effectively monitor people's nutrition data and accurately compute their calorie intake values [10].

One can, therefore, understand and argue that performance and recognition in the world of technology are only attributed to the output but not the input which from this perspective is valid because a software is best recognized to be efficient at how they effectively provide their output. For example, virtual personal assistants, Apple's Siri, Microsoft's Cortana, and Amazon's Alexa, are best recognized not at how they receive data but how they mysteriously process input information and deliver the exact output [2].

In this paper, our priority is to rekindle the recognition of human speech when it comes to providing input data to enable Speech-to-Text systems to render their services in the field of Health, and Education [4, 6, 10]. This is mainly because irrespective of the mysterious processing and computations that come with Speech-to-Text systems, human speech as input data is needed for these processing to commence so they should not under any circumstance be undermined. Therefore, we seek to contribute to Computer science research by highlighting the relevance of human speech when it comes to Natural language processing. We would achieve our priority of rekindling the recognition of human speech by looking at various Speech-to-Text systems and identify how they are dependent on the human voice as input before they can operate. The paper would also address some future work research that should go into Natural Language Processing for the purpose of improvement. These future researches would be unveiled upon discussing the relevance of human speech data in Speech-to-Text systems. Other contributions we would make in this work are as follows:

(1) We provide an educational insight into the Evolution of Natural language processing and how far man has come since the emergence of ELIZA. Understanding the roots of Natural language Processing equips readers with an understanding of the history of Natural language Processing.
(2) Using the insight from some prior research study, we provide evidence on the contribution of Natural Language Processing in today's world.

2 THE EVOLUTION OF NATURAL LANGUAGE PROCESSING

Natural language Processing is a sub-field of Artificial Intelligence that gives a computer program the ability to understand human language as it is spoken. The idea of natural language processing can be attributed to the idea of human communication using natural languages such as English. The intuitive idea is that computers are seeking to process these natural languages used for communication among humans with the sole aim of understanding human speech.

Computers, as we know, are known to only understand machine language and not human language, but as it stands now human-computer communications are possible. The main question is how? ELIZA was an early natural language processing program or a ChatBot created between 1964 to 1966 at the MIT Artificial Intelligence Laboratory by Joseph Weizenbaum [6]. It was recognized as the beginning and the stepping stone for the emergence of Natural language processing. The main aim of ELIZA was to mimic a Rogerian psychotherapist and interact with people on therapy-related issues such as sadness, depression, etc. Speaking of how it does understand and communicates to humans, ELIZA had a list of keywords embedded in a SCRIPT from which it searches within a given input message for their occurrences. Theoretically, what happens is that, when an input message is read by ELIZA, certain keywords which exist in the SCRIPT are searched for in the input message upon which if a keyword is identified, the sentence within which exists the keyword is transformed according to a rule associated with the keyword [6]. The transformation retrieved is then issued out as output by ELIZA. The SCRIPT does not exist in only English but also in Welsh and German as well. To prove its intelligence, ELIZA participated in the Turing test titled the Imitation Game, which was meant to prove machine intelligence [9, 14, 15]. ELIZA passed the test for machine Intelligence due to its effectiveness in mimicking a Rogerian psychotherapist. Despite its marvelous performance, ELIZA had technical problems, some of which include how it would process an input message in the absence of a "keyword" because its source of intelligence and interaction were all regulated from the list of keywords in its SCRIPT. This introductory information about ELIZA was the stepping stone for greater works in the field of Natural language processing. More research was carried out to unveil in-depth understanding in this area of discipline to further help in improving Natural language processing systems through inventions. In the next section, we would briefly look at some advancements made in Natural language processing through inventions to help us understand how far natural language communication between man and machine has come since ELIZA.

2.1 How far natural language communication between man and machine has come since ELIZA

Speaking of the evolution of Natural language processing with ELIZA been its foundation, there has been advancement over generations to improve on natural language processing. Moving from ELIZA, natural language processing today has evolved to also make language to language translations possible. A typical real-life example was the invention of a machine translation system to help bridge the communication gap between patients living in rural areas and doctors in China [12]. This machine conveyed messages about diseases, symptoms, and medical classifications to and from patients

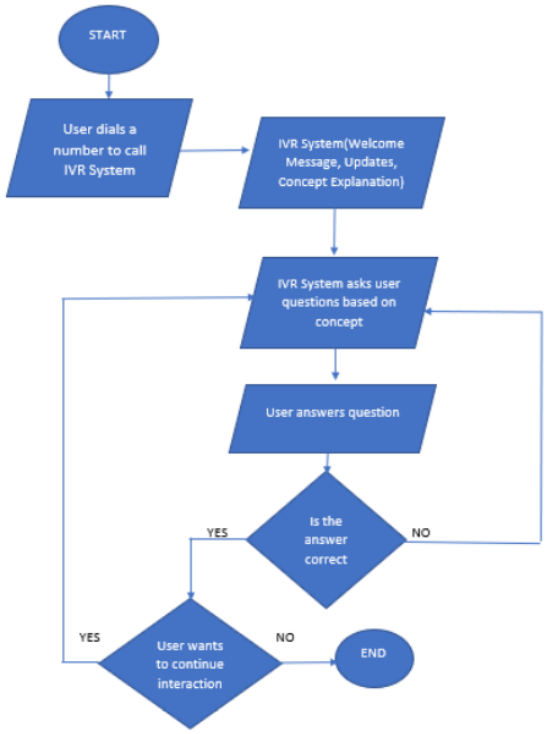
to doctors in languages understood by them. Another popular example is Google Translator that can convert text from one language to another, something which was not present in ELIZA. We also have virtual personal assistants such as Microsoft Cortana, Apple Siri, Amazon Alexa, etc. embedded in portable devices such as phones and personal computers. ELIZA was known to be a stand-alone computer, humans interacted with but in this era, there exist virtual personal assistants with natural language communications techniques to help bridge the gap between AI and human intelligence [11]. These personal assistants have been more portable and widely spread than ELIZA [2].

3 THE RELEVANCE OF HUMAN SPEECH DATA IN SPEECH-TO-TEXT SYSTEMS

Speech-to-Text systems in the course of operation demand human voice as input before they can deliver an output. Therefore without the provision of human voice as input, Speech-to-Text systems cannot function. In this section, we will be looking into some Speech-to-Text systems and understand why such systems cannot operate without the provision of human voice as input.

3.1 The Interactive Voice Response (IVR) System

In the introduction chapter, we looked briefly at two systems, Speech2Health and Nutrition dialogue Speech-to-Text systems which we proved that without human speech their service is worthless in the health sector [7, 10]. In other fields, and to be precise in the Educational sector, speech-to-text systems are also utilized. Let's look at the Interactive Voice Response (IVR) System called Allo Alphabet, a speech-to-text system that aims to assist students to learn at their own pace [4]. Intuitively, what happens is that the user makes a call to a ChatBot that handles the interactions. Upon receiving the call, the IVR System welcomes the user, updates them on their progress, and takes them through the lesson to be taught based on the user's preference. After a lesson, the system plays a pre-recorded audio message with the question and response options(See Fig2 for a detailed interaction of the IVR System). First of all, introducing these activities of the IVR System leaves us with questions such as how is it possible for a system to effectively interact and assist users in such a manner? How are the computations for the IVR System developed? These questions are asked because it marvels us how it is possible. Therefore because of how strange the IVR System functions, much attention is however given to understand how the system works but no attention is given to how human speech as input sets the pace for the IVR System to demonstrate its amazing processes in the first place.



3.1.1 *Fig2: Flowchart Diagram of the IVR System.*

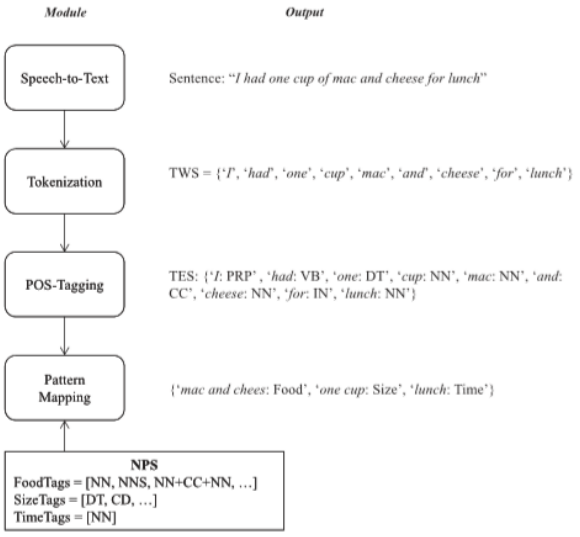
Fig 2 shows schematic diagram of the activities that describe the core intuition of the IVR System.

Now, the attention lies in proving that it was human speech data that set the pace for the IVR System. This proof lies in the function of the IVR System. We identified the function of the IVR System to be a system that assist students in learning at their own pace. So then the proof lies in the question that, if the user had not initiated a call in the first place, informing the IVR System the topic they need help with, would the system have been able to assist them? We can clearly say **"no"** because the system is designed to act based on human input. Therefore, we can say that human speech was the pacesetter for the IVR System to flow.

3.2 Speech2Health System

In this section, we would delve deeper into the Speech2Health system we looked at briefly in the introductory chapter. A high-level overview of the architectural view of the system is illustrated in Fig1. An advanced format in Natural language processing was involved in the makeup of this system. The main components include Speech-to-Text Converter, Nutrient Attribute Extraction, and String Matching [10]. Intuitively, the overview of the system is that the Speech-to-text converter was mainly to help receive audio signals from humans, after which the Nutrient Attribute Extraction would help extract nutrition-related information from human speech input and lastly, the String Matching would help

match the food name from the human speech with predefined food entries in the nutrition database and based on that compute one's calorie intake(see fig3 for more information).



3.2.1 ***Fig3: Detailed activities in Speech2Health System.***
Fig 3 diagram describes the sequence of activities that occur in the Speech2Health system. It shows how after a human speech is received, the method of tokenization breaks the text into individual sets. After this is done, the individual texts are tagged according to their categorizations to enable the Nutrient Attribute extract the food contents from the text. When this is done, the String Matching algorithm is then able to match the food contents with predefined food entries in the nutrition database.

Given such a system, people skip the part of recognizing the relevance of human input data and rather focus much on embracing the system on how it converts human speech into text, how it extracts nutrient data from the text, and lastly, how the string matching is performed. To prove that it was human speech data that set the pace for the Speech2Health System, we can identify that the first and foremost activity from the architectural diagram in Fig1 was the provision of human speech as input. Therefore without the provision of human input in the first place, all the other processes cannot commence. Another way of also understanding that it was human speech data that set the pace for the Speech2Health System is by addressing questions like:

(1) Can the system convert speech to text if no human input speech is given?
(2) Can the system extract Nutrient data from the converted speech to text if no human input speech is given?
(3) Can the system compute one's calorie intake if no human input speech is received for the string matching algorithm to match the food name from the speech with predefined food entries in the nutrition database?

Answering these questions indeed rings the bell that, irrespective of how complex, marvelous, and amazing the Speech2Health system is, without the provision of human speech as input data, it cannot deliver its functions. Speech-to-Text systems in general are not independent systems as they are perceived to be but rather, fully dependent on human speech before they function so, therefore, human speech should also be recognized in Speech-to-Text Systems.

4 THE CONTRIBUTION OF NATURAL LANGUAGE PROCESSING IN TODAY'S WORLD

We have established in the previous section, the relevance of human speech as input data in Speech-to-Text systems and the need for human speech to be recognized in Natural language processing. In this section, we would look into some other contributions achieved in the field of Natural language processing with regards to the collaboration of both human speech and Speech-to-Text Systems.

4.1 Interactive humanoid robots to help children with Autism

Autism is a development disorder that impairs one's ability to communicate and interact. The presence of Interactive humanoid robots today shows the improvement in natural language communication since ELIZA [8]. In today's era, there exist movable humanoid objects called NAO robots that are used in hospitals to assist children suffering from Autism. By assistance, these NAO robots interact with the impaired children in a speech to speech format with the sole aim of helping these children physically and mentally [13]. These robots are also equipped with the capability to blink their eyes, speak, and play music as well to entice their interaction with the impaired children. Prior research shows that these NAO humanoid robots have assisted in improving the communication behavior of children with Autism Spectrum Disorder (A.S.D) [13].

4.2 Intelligent Personal Assistants(I.P.As) to assist the Elderly

Moving over to the social sector, natural language processing has advanced to become more portable. In this dispensation, we can see the usefulness of Amazon Alexa, Google Assistant, Microsoft Cortana, and Apple Siri in our everyday life. Specifically, they have also been assisting the elderly who most of the time are socially isolated. This is because the elderly who are been isolated are more likely to experience mental disorders like depression [1]. Therefore the usage of these Personal assistants to keep them company is proving to help engage the elderly effectively thereby preventing the occurrence of depression [1].

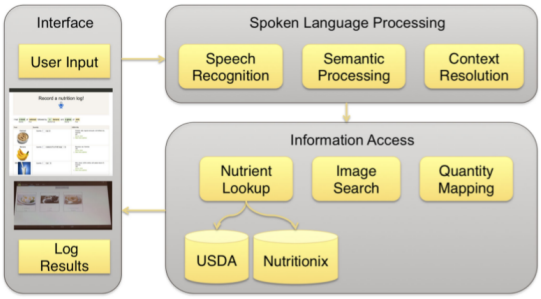
4.3 Interfacing of Artificial Intelligence with Computer Algebra Systems(C.A.S)

Another advancement in the field of Education is the inclusion of Artificial intelligence to help in mathematical symbolic problem-solving. Symbolic reasoning was a crucial

task for many scientists, engineers, and mathematicians to assist them in performing scientific and algebraic computations. Computer algebra systems(C.A.S) were in use to help scientists and mathematicians but because the complexity of models and the size of data sets used in these fields grew at an increasing pace, there seem to be irrelevant [5]. Due to the introduction of Artificial Intelligence, these computer algebra systems are embedded with AI-related techniques to enable scientists to perform creative symbolic problem solving with the computer algebra systems [3, 5]. Another benefit that comes with Computer algebra system interfacing with AI is that creative problem solvers now can define their notations as they work.

4.4 Nutritional Dialogue System to educate patients concerning nutrition

This Natural Language Processing system is very similar to the Speech2Health System. The Nutritional dialogue system is also a Speech-to-Text System that includes Convolutional Neural Networks (C.N.N) in its processes [7]. This advanced Speech-to-Text contributes massively in assisting users by establishing a spoken dialogue platform to engage users in health-wise conversations after receiving their recorded meal information(see Fig 4). The System can also initiate an interaction session with users to advise them on their nutritional contents and ask follow-up clarification questions about user's health concerns.



4.4.1 *Fig4: Activities in Nutritional Dialogue System.*

Fig 4 shows the schematic representation of the Nutritional Dialogue System. The diagram goes in-depth to illustrate the process from where the user records their meal information, followed by a spoken language understanding by the dialogue system, nutrient database lookup, and finally responding to the user with the results.

5 FUTURE WORK

For future work, research should be conducted to mainly focus on explaining the major processes that go into Speech-to-Text Systems. This is because the majority of insights given are Computer Science-oriented therefore making it difficult for non-computer science persons to understand. Further research should also be conducted on measuring the impact of N.L.P Technology by accessing its usability and accessibility. This would help to understand areas in Natural language processing that need re-construction as well as improving the interfaces of these Speech-to-Text systems such that they become more user-friendly.

6 CONCLUSION

In this paper, we provided educational insight into the Evolution of Natural Language Processing by looking at the activities of ELIZA, a ChatBot created between 1964 and 1966. We also delved deeper to create awareness about how we have transitioned with regards to Natural Language Processing from ELIZA to this present day by looking at some advancements in the Social, Technological, Educational, and Health sectors. We also identified the need to recognize human voice input in Speech-to-Text systems by looking at two diverse systems namely the Interactive Voice Response (IVR) System and the Speech2Health System and indicating how relevant it is for human speech input to be provided before these Speech-to-Text Systems can commence operations. We further unveiled some amazing contributions of Natural Language Processing in this dispensation in diverse sectors. Finally, Speech-to-Text systems, therefore, are efficient systems not just because of their vast amazing processes and computations, but also because of the contribution of human speech as input data to help commence its operations [4, 10].

REFERENCES

[1] João Barroso Dennis Paulino Maria João Monteiro Hugo Paredes Arsénio Reis, Isabel Barroso and Vitor Rodrigues. 2018. Using intelligent personal assistants to assist the elderlies. *2018 2nd International Conference on Technology and Innovation in Sports, Health and Wellbeing (TISHW)* (2018), 1–5.

[2] Gamal Bohouta and Veton Z Këpuska. 2018. Next-generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home). *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)* (Jan. 2018), 99–103.

[3] Guijie Li Hang Zhao and Wei Feng. 2018. Research on Application of Artificial Intelligence in Medical Education. *2018 International Conference on Engineering Simulation and Intelligent Control (ESAIC)* (2018), 340–342. https://doi.org/10.1109/ESAIC.2018.00085

[4] Kaja Jasinska and Amy Ogan. 2019. You Give a Little of Yourself: Family Support for Children's Use of an IVR Literacy System. *ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS) (COMPASS '19)* (2019), 86–98. https://doi.org/10.1145/3314344.3332504

[5] Benjamin T. Jones. 2018. Human-AI Interaction in Symbolic Problem Solving. *2018 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)* (2018), 265–266.

[6] Weizenbaum Joseph. 1966. ELIZA—a Computer Program for the Study of Natural Language Communication between Man and Machine. *Commun. ACM* 9, 1 (Jan. 1966), 36–45. https://doi.org/10.1145/365153.365168

[7] Mandy Korpusik and James Glass. 2017. Spoken Language Understanding for a Nutrition Dialogue System. *IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING* 25, 7 (July 2017), 1450–1461. https://doi.org/10.1109/TASLP.2017.2694699

[8] Tzu-Chien Liu and Maiga Chang. 2008. Human-Robot Interaction Research Issues of Educational Robots. *2008 Second IEEE International Conference on Digital Game and Intelligent Toy Enhanced Learning* (2008), 209–210. https://doi.org/10.1109/DIGITEL.2008.31

[9] Guisheng Chen Yuchao Liu Liwei Huang, Haisu Zhang and Deyi Li. 2017. FROM TURING MACHINE INTELLIGENCE TO COLLECTIVE INTELLIGENCE. *2012 IEEE 2nd International Conference on Cloud Computing and Intelligence Systems* 3 (2017), 1171–1177.

[10] Sepideh Mazrouee Niloofar Hezarjaribi and Hassan Ghasemzadeh. 2018. Speech2Health: A Mobile Framework for Monitoring Dietary Composition From Spoken Data. *IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS* 22, 1 (Jan. 2018), 252–264.

[11] Narasimhan R. 1990. Human intelligence and AI: how close are we to bridging the gap? (open line). *IEEE Expert* 5, 2 (1990), 77–79.

[12] Su Sha Rean Aierken, Li Xiao and Dawa Yidemucao. 2016. Multiple-Language Translation System Focusing on Long-Distance Medical and Outpatient Services.

International Journal of Advanced Research in Artificial Intelligence 5, 4 (2016), 471–475.

[13] Luthffi Idzhar Ismail Salina Mohamed Fazah Akhtar Hanapiah Syamimi Shamsuddin, Hanafiah Yussof and Nur Ismarrubie Zahari. 2012. Humanoid Robot NAO Interacting with Autistic Children of Moderately Impaired Intelligence to Augment Communication Skills. *Procedia Engineering* 41, 1 (2012), 1533–1538. https://doi.org/10.1016/j.proeng.2012.07.346

[14] Alan M. Turing. 1950. COMPUTING MACHINERY AND INTELLIGENCE. *Mind* LIX, 236 (Oct. 1950), 433–460. https://doi.org/10.1093/mind/LIX.236.433

[15] Kevin Warwick and Huma Shah. 2014. Good Machine Performance in Turing's Imitation Game. *IEEE Transactions on Computational Intelligence and AI in Games* 6, 3 (Sept. 2014), 289–299. https://doi.org/10.1109/TCIAIG.2013.2283538