Examining the Impact of Data Type and Structure on Speech-Based Accessible Computing

Andrew DiStasi
Rochester Institute of Technology
Rochester, NY
add5980@rit.edu

1. ABSTRACT

Continual Improvement in the fields of Speech Synthesis and Natural Language Processing have given rise to an increase of applications and tasks enhanced by the use of Dialog Systems and Intelligent Personal Assistants. These improvements could have a substantial impact on the ability of developers to utilize speech-based technology to create a more accessible technology experience for blind users and physically disabled users. Traditionally, the greatest roadblock to the acceptance of speech-based systems and applications has been twofold – First, it is difficult for computers to reliably generate utterances that sound natural and appropriately conversational, and second, that applications can struggle with effectively relaying information to the user through the medium of speech utterances (especially when compared to some form of Graphical User Interface. A potential means of improving the quality of speech-based systems in these regards is to examine the appropriateness of different data types – structured data, semi-structured data, and unstructured data – for conveying utterances to a user. To that end, I plan to develop a speech-based IPA-Style application that will guide a user through a recipe utilizing each of the data types to allow for evaluation and comparison of the effectiveness of each data type.

2. INTRODUCTION

While the traditional means of input for computer systems – some form of keyboard and mouse – has long been the standard, it alienates multiple user segments in varying capacities, including vision-impaired users, elderly users, physically disabled users, and even users whose hands are just otherwise occupied. Natural Language (Speech) is an intuitive modality of interaction and data transferal most users are familiar with, making it uniquely qualified to serve as a method of interaction with a computer system. While many speech technologies exist today, these implementations are not sufficient to call these accessibility issues "solved." Further examination of the appropriateness of speech to convey different types of data is necessary to identify and support the best practices and areas of focus for developing alternate speech-based means of accessibility, particularly in the realm of Intelligent Personal Assistant-style applications in mobile devices.

Recent trends have seen an increased focus on accessible computing. As the technology has begun to catch up with demands posed by various types of disabilities, developers have increased their efforts to broaden the range of how users can interact with computer systems. Historically, devices such as text-to-speech (TTS) systems, magnifying screens, and screen readers have helped users overcome vision-based disabilities to continue utilizing computers. As speech-based technology, continues to mature, however, it is imperative that we, as developers, continually strive to improve these tools for making computing accessible. One of the imperatives of accessibility is that computing should be made available to all users, regardless of ability. While TTS Systems and other speech-based technologies certainly exist, they haven't proved intuitive enough to warrant wide-scale adoption as a main mode of computer interaction. Additionally, many of the commonly found examples of TTS simply echo alternatively defined text from an application (Screen Readers) or pass an inputted string of speech to a predefined application or search engine (Intelligent Personal Assistants).

Part of this can be attributed to some of the characteristics of speech that make it markedly different from other standard methods of input, especially visual input. Speech as a means of communicating information can be defined as transient, invisible, and asymmetric [4]. It is transient because it doesn't ever maintain any state – once it has been spoken it is gone. It is invisible because there is never any non-audial representation of it, nothing can be seen. Finally, it is asymmetric because users tend to be able to speak or generate information far more efficiently than they can hear or process information. As a result of these issues, many speech-based applications do not feel natural or intuitive to users, which can be one of the largest impediments to widespread adoption of a system or means of input. Therefore, in order to increase the usability, and potentially adoption rate, of speech based technology, we need to work to make these systems feel more natural. One potential area of study that could increase the effectiveness of these applications is a study into the effectiveness of different types of data in conveying information between the system and a user.

There are three main types of data that can be used to transmit information between the system and the user – Structured Data, Semi-structured Data, and Unstructured Data. Structured Data is, essentially, highly organized information that has been sufficiently detailed and annotated such that its inclusion in a relational database or search engine would be seamless. It can frequently be displayed in a tabular format. Unstructured Data is the type of data that will likely feel most natural to users, because it is essentially natural text. This data type doesn't have any explicit organization, structure, or annotation, it is simply human language transcribed into text. Common examples can be found on any social media site or internet forum where users are able to publish posts written in their own natural language. Semi-structured data takes elements of both structured and unstructured data, combining them in a hybrid approach. It doesn't conform to any formal structure associated with relational databases or other forms of organization used in structured data, but it does contain tags or other markers to contain semantic elements and enforce hierarchical structure between different data points. This format has become increasingly prevalent with the growth of the internet, where data has been represented and stored in different formats. This data format is extremely common on the Internet and on web applications, with common examples of this format including HTML, XML, and JSON.

Each of these data structures offers different strengths and weaknesses with regard to speech technologies. Unstructured data is by far the most natural sounding and intuitive to a user, but proves the most challenging for a computer to parse and understand. Conversely, Structured data integrates seamlessly into relational databases and computers, but can be hard for users to understand and doesn't lend itself well to natural text. Semi-structured text again finds a middle ground between the two, generally containing some representation of natural language, with some small associated set of tags or hierarchical structure. These strengths and weaknesses have a clear correlation to effectiveness in TTS applications – unstructured data makes the application feel intuitive and natural for a user, but structured data makes it easier for the system to process inputted information. To develop truly effective speech-based software, it is important that we find which data type best transfers data between the user and system while maintaining a degree of naturalness for the user.

To that end, this study will be focused within the domain of recipes, as the task of preparing a recipe is suitable for speech-based guidance and recipe data exists in all three data formats. Wasinger states that speech is an appropriate form of input when no keyboard is available, the task makes use of deep and complex menu structures, and the user's hands and eyes are already occupied. These are all conditions that are met in the process of cooking, so we can safely conclude this domain is appropriate for speech-based technology [4]. Additionally, recipe data exists in all three data formats, allowing for comparison of each format's effectiveness across one domain. Recipes are commonly found in semi-structured and unstructured format. A recipe that has a specific list of ingredients, utensils required, or metadata about the recipe (such as time requirement, serving amount, etc.) meets the requirements of the semi-structured format. A recipe in the unstructured format would consist only of textual instructions on how to complete the recipe. Recipes in a structured data format are a little more uncommon, but there are several online datasets that curate structured recipe data. Google has a collection of "Rich cards" of recipe data stored in a JSON-LD format that integrates with their search engine.

3. RESEARCH OBJECTIVE

The aim of this Research Project is to identify which data type – unstructured, semi-structured, or structured – is the most effective for use in Intelligent Personal Assistant style applications and in speech-based accessible computing. Additionally, this project will aim to develop and evaluate a prototype of such an application for the purposes of evaluating how the usability and effectiveness of the application changes when using each different data type within the scope of one sample domain.

4. RESEARCH QUESTIONS

This Research Project aims to identify if the structure of data used in a speech-based application has any measurable effect on the overall effectiveness and accessibility of the application. Additionally, it attempts to identify if the type of data used impacts a user's ability to understand and comprehend the information relayed to them. Finally, it attempts to answer which, if any, of the three data types is best suited for use in Intelligent Personal Assistant style applications and speech-based accessible computing.

5. LITERATURE REVIEW

The fields of Speech Recognition, Natural Language Processing, and Natural Language Generation have seen a recent surge in attention as the relevant technology becomes more and more sophisticated and the demand for improved accessibility increases. One distinction that is important to make is the distinction between Natural Language Processing and Natural Language Generation. Natural Language Processing is, essentially, the ability of a computer program to understand human language as it is spoken. Alonso et al. defines Natural Language Generation as a field that "uses analytics, AI, and NLP to obtain relevant information about non-linguistic data and to generate textual summaries and explanations of these data which help people understand and benefit from them" [1]. Alonso et al. conducted a survey of several papers devoted to "recent and prominent developments in the field of NLG with Computational Intelligence" [1], examining how Computational Intelligence and Soft Computing allow NLG Systems to represent and deal with the inherent imprecisions and uncertainties of human language. They explore the benefits of combining CI and NLG approaches to improve the handling of vague tasks within the NLG Pipeline and to help developers quickly adapt NLG systems to a new domain.

Another field being explored in conjunction with NLG is that of Linguistic Description of Data (LDD). Ramos-Soto et al. explore the current state of the Natural Language Generation and Linguistic Description of Data fields, examine how these fields generate easily understandable information from data, and explore "potential points of mutual interest and convergence between both fields" [7]. They explore both fields across multiple applications within the meteorology domain, using it as a means to examine the current state of each field before turning to potential areas of intersection between the two. Ramos-Soto et al. posits that "deeper insight into NLG will greatly benefit LDD researchers, especially regarding the development of applied approaches for practical problems" and that "LDD can be a field of interest for NLG researchers in several respects, including quantified sentences and potential derived extensions, evaluation criteria, algorithms, and more importantly, imprecision handling" [7]. These potential collaborations set the stage for a more complete application that addresses some weaknesses and improves upon some strengths found within applications in these two domains. This collaboration is relevant for speech-based accessible technology, because both fields can play a prominent role in improving the functionality and naturality of these applications by allowing applications to more completely annotate and understand the data that is passed in.

Kumagai et al. conducted another exploration into Natural Language Generation, this time focusing on employing a Monte Carlo Tree Search to account for situational nature of the structure and content of human speech. Specifically, they "build a search tree of possible syntactic trees to generate a sentence, by selecting proper rules through numerous random simulations of possible yields" [2]. The primary means of evaluation for this work involves giving an NLG machine the ability to effectively evaluate whether or not a constructed sentence is natural. To achieve this, Kumagai et al. utilize a dual-method evaluation score, combining an evaluation of syntactic structure and of the n-gram language model. These two results can be combined to "score" a system on pass/fail criteria to determine if the sentence is natural. This ability offers a key component of functionality to a speech based system, as how natural using it feels is a large indicator of the system's potential acceptance by users. A pass/fail evaluation of how natural an utterance is could be a vital part of the evaluation process for speech-based software, as accessibility tends to be one of the largest gates to widespread acceptance.

Now that speech and Natural Language Generation have been examined, we must examine the other relevant component: Accessibility. As software prevalence becomes denser every day and human life expectancy increases, more individuals are using technology and software more frequently, for longer. "Accessible software does not simply mean that a person with some impairment will be able to use it, it means much more. With accessible software, people who were not able to do simple everyday things, which make up the daily life of much of the population, are now able to accomplish them" [10]. Silva et al. explore a plan for improving the awareness of relevant accessibility issues and implementation, and exploring the relationship between "relevant accessibility documentation and its appropriate type of user interface" [10]. They motivate their research with a cursory search through two paper databases: The Web of Science and Science Direct. Searches of the phrases "Web Accessibility" and "Software Accessibility" show that approximately 25 papers were related to these topics out of millions of stored papers. While not directly correlated to any component of speech or a particular accessible system, establishing a shared understanding of the common issues that accessibility solutions must answer is a necessary component of identifying the best practices for accessible systems.

Now that each component has been explored, we can begin looking at some applications of these ideas into a practical domain. Lacey et al. and Norman et al. both explore NLG and NLP technology in generating human readable result overviews from differing types of clinical data. Lacey et al. focused on utilizing natural text descriptions of doctor observations in Epilepsy clinic letters to "extract meaningful and technically correct clinical information from free text sources" [3]. Making use of IBM Watson Content Analytics software (ICA), they defined annotations based on language characteristics to create parsing rules and an NLP pipeline that highlighted and extracted relevant items from clinic letters, including "symptoms and diagnoses, medication and test results, as well as personal identifiers" [3]. A series of epilepsy clinic letters, containing a mix of "new patient" letters and "follow-up" letters across 12 different doctors, were anonymized and fed into the ICA system. Lacey et al. focused on extracting a discretized epilepsy type, cause, age of onset, medical test results, prescribed medication, and clinic date. Their results show a startlingly high accuracy for all extracted features (the lowest of which was 95%), indicating that, at least in this limited domain example, ICA is capable of properly extracting information from unstructured text.

Norman et al. explore a slightly different type of data, attempting to extract a Pediatric Appendicitis Score (PAS) from a combination of structured and unstructured data. The Pediatric Appendicitis Score is used to aid physicians by automatically generating a score ranking the likelihood that a pediatric patient had appendicitis. This is valuable because the harmful effects of excessive exposure to radiation dictate that diagnostic imaging be minimized, especially in child patients [6]. To that end, a PAS Score below 4 indicates that there is a low suspicion for appendicitis, meaning that imaging is not required unless additional symptoms present themselves. A score above 8 also removes the need for imaging, as this score should automatically lead to a surgery consultation because the likelihood of appendicitis is very high. Norman et al. created a software application that performed NLP preprocessing and feature extraction on a set of text before feeding that information into a model that utilizes a series of classifiers to extract and tag relevant textual data [6]. They found that a Logistic Regression classifier gave them an F-score of 0.9874, indicating that it was very effective at correctly extracting and classifying PAS data from both structured and unstructured tests.

Another experiment conducted by Sauer et al. focused on a combination of structured and semi-structured (rather than unstructured, like the previous two papers) data. Sauer et al. utilized an NLP Tool to extract Pulmonary Function Test (PFT) Reports from Veteran Affairs data of these types. These PFTs are "objective estimates of lung function, but are not reliably stored within the Veteran Health Affairs data systems as structured data" [8]. Data was extracted from the reports of patients at seven VA Medical Centers who suffered from asthma and was fed into a NLP tool Sauer et al. developed. Performance was judged against a human reference standard over 1,001 randomly sampled documents. They found that the tool demonstrated a precision of 98.9% in the validation set, indicating that it can observably improve identification of PFTs in medical research and treatment. However, Sauer et al. caution that it would be erroneous to assume that "a single domain of clinical data can completely capture the scope of a disease, treatment, or clinical test" [8].

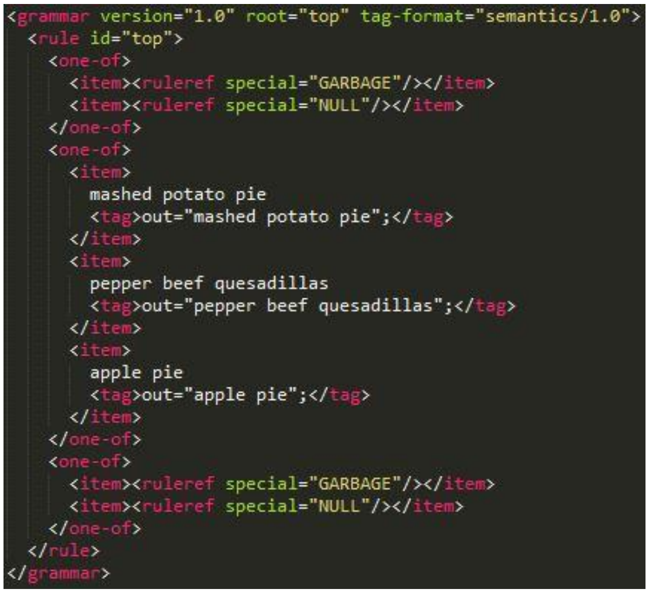
The final healthcare related study collected for this paper focuses on using an NLP tool for large-scale data extraction from Echocardiography Reports. Nath et al. observed that because Echocardiography is one of the most commonly ordered diagnostic tests in cardiology, "large volumes of data are continuously generated from clinical notes and diagnostic studies catalogued in electronic health records (EHRs)" [5]. One of the major barriers to leveraging this unstructured data to improve the quality of care for patients is that there are few viable tools that allow accurate extraction of high-quality data from such a large volume of various forms of unstructured data [5]. To that end, Nath et al developed an NLP tool called EchoInfer, that allows for automatic extraction of "data pertaining to cardiovascular structure and function from heterogeneously formatted echocardiographic data sources" [5]. Data elements were extracted and structured into various data formats before being preprocessed and having a series of document and sentence segmentations performed to isolate text relating to certain features and generate relationships between these sentences and the relevant features. Nath et al. analyzed 15,116 echocardiography reports from 1,684 patients, extracting 59 quantitative and 21 qualitative data elements per report [5]. EchoInfer achieved a precision of 94.06%, a recall of 92.21%, and an F1-Score of 93.21% across all data elements in a test subset of 50 reports. Given these results, we can conclude that EchoInfer's NLP Processing permits large-scale extraction across various data types pertaining to echocardiographic reports with a high degree of precision, accuracy, and recall.

To explore a domain outside of healthcare, we turn to research performed by Schlunz et al., where they examine the accessibility in TTS synthesis for South African Languages. They examine three use cases where multilingual individuals using some form of Augmentative and alternative communication were observed to measure a "baseline integration of the existing Qfrency TTS voices into a selected AAC system and to evaluate the user experience" [9]. Grid 3, an AAC system sold and commonly used in South Africa was integrated with the Qfrency TTS voices and customized to build text interfaces with simple South African sentences. Literate AAC users were recruited to perform acceptance testing on this new tool, focusing on how natural and intelligible the TTS voices were when using the application. Users were asked to utilize closed-form answering machines to rank the prosody, pronunciation, and intelligibility of the system. Intelligibility was scored high fairly consistently, but naturalness ratings were "more spread out between the two poles of robotic and human-like synthetic speech" [9]. This shows that while current technology is capable of making systems that can be understood, there are still steps to be taken to improve how natural utilizing such a system feels. Additionally, none of the previous studies focused on the appropriateness of a given data type as a variable in the success or failure of the application. This motivates a line of inquiry into determining how the use of these different data types could affect the naturalness of NLP, NLG, and TTS Systems. All of these studies serve as proofs of concept for applications that accept user data (in various formats, including structured, semi-structured, and unstructured data) and generate some form of natural text to display to the user.

6. RESEARCH PLAN

To properly examine the impact of different data types and formats on the effectiveness of speech-based accessible applications, I will develop an application designed to use speech technology to guide users through the process of preparing a recipe. This application will be based off a similar application I developed called GORDON – Gourmet Oral Recipe Dictation Or Narration System – for research with Dr. Emily Prud'hommeaux and the Rochester Institute of Technology Language Science Department.

The application will be a web application based in VoiceXML, a simple programming language designed to create voice applications in a web environment. See Figure 6.1 for a sample of VoiceXML code. VoiceXML is well suited for the initial development of this application because it accommodates the creation of the multi-leveled and complex datasets necessary to convey recipe information to a user, but does so in a simple and easy to understand XML-like structure. Additionally, the ability to run this as a web application minimizes set up or configuration work for the application should multiple machines be used in the development and evaluation process.



```
<grammar version="1.0" root="top" tag-format="semantics/1.0">
  <rule id="top">
    <one-of>
      <item><ruleref special="GARBAGE"/></item>
      <item><ruleref special="NULL"/></item>
    </one-of>
    <one-of>
      <item>
        mashed potato pie
        <tag>out="mashed potato pie";</tag>
      </item>
      <item>
        pepper beef quesadillas
        <tag>out="pepper beef quesadillas";</tag>
      </item>
      <item>
        apple pie
        <tag>out="apple pie";</tag>
      </item>
    </one-of>
    <one-of>
      <item><ruleref special="GARBAGE"/></item>
      <item><ruleref special="NULL"/></item>
    </one-of>
  </rule>
</grammar>
```

Figure 6.1 – VoiceXML Code used to attach GARBAGE & NULL rulerefs, allowing for more natural input.

Upon completion of development for the application, thirty sighted individuals of varying technological familiarity, age, and cooking ability will be selected to prepare one recipe twice – once using the application and once using a traditional cookbook. These individuals will be divided into 3 three groups of ten. Ten of the users will complete a series of tests with a version of the application that returns recipe information in a structured format. Another ten users will complete a series of tests with a version of the application that returns recipe information in a semi-structured format. The final ten users will complete a series of tests with a version of the application that returns recipe information in an unstructured format. Half of the users in each group will complete the tests using the application first and half the users in each group will complete the tests using a traditional cookbook first. This helps account for users becoming familiar with the recipe after the first trial, which could lead to faster and more confidant preparation of the recipe. Thirty participants were selected for this experiment because it allows 5 data points for each possible configuration (application first and cookbook first for each of the three data types), which is close to the number of initial evaluations done in similar studies.

Statistics from both trials for each user will be recorded, and users will be asked to provide a series of ratings and feedback after the completion of both trials. The specific statistics and feedback received will be detailed in the Evaluation Plan section. These results will be used to identify strengths and weaknesses of each data type and to refine the application and information that is returned to the user to better respond to user needs. Once this has been completed, the feedback generated by the application for each data type will be separately uploaded to Amazon Mechanical Turk for evaluation on a series of metrics that will be discussed in the evaluation plan section. This is done to generate as large of a dataset as possible to minimize the impact of any individual tendencies of users. These results will be evaluated for statistical significance and used as the primary metric to determine if one of the data types offers a superior performance to the others.

7. EVALUATION PLAN

In order to effectively evaluate the results of the user tests performed on the application created as part of the processes discussed in section six, a series of metrics about the trials and user feedback, both quantitative and qualitative, will be collected. For data about the trials themselves, we will track the time spent completing each individual step, the total time spent completing the recipe, the number of times the user re-referenced the cookbook or the application, and the number of times the application failed to correctly capture user input. Additionally, the user will be asked to provide quantitative rankings of their comfort level with the application, how natural they perceived the utterances generated by the system to be, and how effectively they felt they were able to input and receive information from the system. Finally, users will be asked to offer any qualitative feedback they may have on the system or the process of interacting with it. The qualitative results will be averaged across the ten users for each data type and compared and examined for statistical significance. These results, along with qualitative user feedback will be used to fine-tune the system.

The second set of evaluations completed will include an analysis of the application-generated feedback and responses by users on Amazon Mechanical Turk. Users will be asked to listen to audio files containing the utterances generated by the application and will be asked to quantitatively rate how natural they consider these utterances to be and how effective they felt the utterances were at conveying information about the recipe. This data will be accumulated and examined for statistical significance to determine if any correlation exists between the naturality of the utterances and the data type used or the effectiveness of the utterances and the data type used. If applicable, these ratings will be used to identify which, if any, of the data types offers superior performance over the others for use in speech-based accessible applications.

8. REFERENCES

[1] Alonso, Hose M, and Alberto Bugarin. "Natural Language Generation with Computational Intelligence." *IEEE Xplore*, Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS), 19 July 2017, ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7983468.

[2] Kumagai, Kaori, et al. "Human-like Natural Language Generation Using Monte Carlo Tree Search." *Proceedings of the INLG 2016 Workshop on Computational Creativity and Natural Language Generation*, no. 2016, Sept. 2016, pp. 11–18., www.aclweb.org/anthology/W16-5502. Association for Computational Linguistics.Human-like Natural Language Generation Using Monte Carlo Tree Search

[3] Lacey, Arron S., et al. "Obtaining structured clinical data from unstructured data using natural language processing software." *International Journal of Population Data Science*, Aug. 2016, pp. 1–2., ijpds.org/article/view/381/362. IJPDS.

[4] Wasinger, R.: Dialog-based user interfaces featuring a home cooking assistant, University of Sydney, Australia (2001) (unpublished manuscript). http://rainet.wasinet.com/hca/4%20Page%20Paper.pdf

[5] Nath C, Albaghdadi MS, Jonnalagadda SR (2016) A Natural Language Processing Tool for Large-Scale Data Extraction from Echocardiography Reports. PLOS ONE 11(4): e0153749. https://doi.org/10.1371/journal.pone.0153749

[6] Norman, Brittany, et al. (2017). Automated identification of pediatric appendicitis score in emergency department notes using natural language processing. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7897310

[7] Ramos-Soto, A., et al. "On the role of linguistic descriptions of data in the building of natural language generation systems." *Fuzzy Sets and Systems*, vol. 285, 15 Feb. 2016, pp. 31–51. *Research Center on Information Technologies*, (CiTIUS), www.sciencedirect.com/science/article/pii/S0165011415003085.

[8] Sauer, Brian C. et al. "Performance of a Natural Language Processing (NLP) Tool to Extract Pulmonary Function Test (PFT) Reports from Structured and Semistructured Veteran Affairs (VA) Data." *eGEMs* 4.1 (2016): 1217. *PMC*. Web. 26 Oct. 2017.

[9] Georg I. Schlünz et al. 2017. Applications in accessibility of text-to-speech synthesis for South African languages: initial system integration and user engagement. In *Proceedings of the South African Institute of Computer Scientists and Information Technologists* (SAICSIT '17). ACM, New York, NY, USA, Article 32, 10 pages. https://dl.acm.org/citation.cfm?id=3129445

[10] de Sousa e Silva J., et al. (2017) Making Software Accessible, but not Assistive: A Proposal for a First Insight for Students. In: Rocha Á., Correia A., Adeli H., Reis L., Costanzo S. (eds) *Recent Advances in Information Systems and Technologies*. WorldCIST 2017. Advances in Intelligent Systems and Computing, vol 570. Springer, Cham