B. Comp. Dissertation

**Talk-to-Code: Coding by Dictation**

By

Lam Zhen Zong Nicholas

Department of Computer Science

School of Computing

National University of Singapore

2017/2018

B. Comp. Dissertation

**Talk-to-Code: Coding by Dictation**

By

Lam Zhen Zong Nicholas

Department of Computer Science

School of Computing

National University of Singapore

2017/2018

Project No: H0201280

Project Supervisor: Assoc Prof. Ooi Wei Tsang

Deliverables:

Report: 1 Volume

Contents

[1 Introduction 1](#_Toc497680348)

[1.1 Project Objectives 1](#_Toc497680349)

[2 Literature Review 1](#_Toc497680350)

[3 Progress made so far 1](#_Toc497680351)

[3.1 Project Overview 1](#_Toc497680352)

[3.2 Speech Recognizer Module 3](#_Toc497680353)

[3.2.1 Experiment with Basket of keywords 3](#_Toc497680354)

[3.2.2 Experiment with script for a sample program 3](#_Toc497680355)

[3.3 Structured Language 4](#_Toc497680356)

[3.4 Word Corrector Module 6](#_Toc497680357)

[3.4.1 Word Corrector Phase One 7](#_Toc497680358)

[3.4.2 Word Corrector Phase Two 7](#_Toc497680359)

[3.5 Word Parser Module 7](#_Toc497680360)

[3.6 Structured Command Parser Module 8](#_Toc497680361)

[4 Plan for the next semester 8](#_Toc497680362)

[4.1 Improving the Structured Language 8](#_Toc497680363)

[4.2 Adding more constructs to the Structured Language 9](#_Toc497680364)

[4.3 Improving Word Corrector Module 9](#_Toc497680365)

[4.4 Improving overall program and feedback 9](#_Toc497680366)

[5 References 9](#_Toc497680367)

# 1 Introduction

## 1.1 Project Objectives

Coding by Dictation is a continuation of the previous project ‘Talk-to-Code: From Structured Command to Source Code’ (Gao, 2016) which creates a foundation for creating a hands-free natural language programming application. This application allows users to write program code by just dictating the program to the computer via voice input.

This project aims to enable people with disabilities or conditions such as the Carpal Tunnel Syndrome (CTS) to write computer programs. CTS (also known as Repetitive Strain Injury) is a medical condition which affects the median nerves of the hand, causing pain and discomfort to the user. It is found that majority of people diagnosed with CTS tend to perform repetitive tasks of some sort involving the use of fingers or hands. (Crouch, 1995) Thus, it is of no wonder that programmers fall under this category, as they perform repetitive typing tasks to write computer programs.

This project will help programmers to write program code without using their hands, therefore reducing the risk of contracting CTS, or even enable CTS patients to write code once again.

# 2 Literature Review

# 3 Progress made so far

## 3.1 Project Overview

This project focuses on creating an end-to-end program which converts a user’s voice into C program code. The ideal situation would be to have a program which allows programmers to speak in a human and natural way, and translates that to code subsequently. However, it is noted that there can be many different ways to verbalize the same piece of code, and that natural English commands were rather ambiguous in the sense that there can be many different interpretations of that natural English command. (Chew, 2016)

For instance, in order to declare an integer variable with value zero, one programmer might verbalize “Create integer variable with value zero” while another programmer might say it as “Declare an integer variable equal zero”. Both programmers are trying to write the exact same piece of code through different ways of saying it, but it is not easy to have a program that is able to decipher and know that these two different verbalizations should result in the generation of the exact same piece of code.

As for the latter case regarding the ambiguity of the natural English commands, the English command “declare integer x, y, z” can be interpreted (in code) as “int xyz;” or “int x, y, z;”. This ambiguity can result in the 2 different interpretations as outlined above, and this causes problems as the program will not know which piece of code to generate and what the user actually intends to write.

In order to eliminate ambiguities, this project will introduce a structured and fixed way of expressing programming constructs, while incorporating some natural English in my structured language as well. The structured language will be presented in a later section in this paper. The system architecture of the Coding by Dictation application at the current stage is depicted in the diagram below.

Voice input from user

Structured Command Parser Module

*Figure 1: System Architecture of Coding by Dictation application*

Structured Command

Processed text

English text (structured language)

Program Code

Word Parser Module

Word Corrector Module

Speech Recognizer Module

## 3.2 Speech Recognizer Module

My Speech Recognizer Module converts voice input from the user into English text, using the python Speech Recognition library (Zhang, 2017). The python Speech Recognition library supports several APIs in which we can use to perform the speech recognition. Out of the several APIs, I have picked 3 of them (mainly Google Speech Recognition, Google Cloud Speech API, and Microsoft Bing Voice Recognition) to carry out further testing and experiment before deciding on one of them to use for my Coding by Dictation application.

### 3.2.1 Experiment with Basket of keywords

After coming up with an initial draft of my own structured language, I have picked out some of the keywords and conducted an experiment with 10 subjects. The subjects were asked to record their voice into an audio file which consists of the keywords. These audio files are then passed into my speech recognizer module with the use of 3 different APIs as mentioned above. The summarized results of the experiment are shown below in Figure 2 (those in yellow signifies the highest scoring API).

*Figure 2: Summarized results of Speech Recognizer module with basket of keywords*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Keywords** | **Google Speech Recognition** | **Google Cloud Speech API** | **Microsoft Bing Voice Recognition** | **Total** |
| Equal | 7 / 10 | 9 / 10 | 6 / 10 | 22 / 30 |
| If Then Else | 5 / 10 | 5 / 10 | 5 / 10 | 15 / 30 |
| End If | 0 / 10 | 5 / 10 | 0 / 10 | 5 / 30 |
| Declare integer | 5 / 10 | 7 / 10 | 0 / 10 | 12 / 30 |
| Size | 6 / 10 | 9 / 10 | 8 / 10 | 23 / 30 |
| Index | 9 / 10 | 10 / 10 | 10 / 10 | 29 / 30 |
| Create Function | 4 / 10 | 9 / 10 | 6 / 10 | 19 / 30 |
| Return Type | 1 / 10 | 8 / 10 | 2 / 10 | 11 / 30 |
| Parameter | 3 / 10 | 9 / 10 | 1 / 10 | 13 / 30 |
| Call | 5 / 10 | 5 / 10 | 10 / 10 | 20 / 30 |
| For | 0 / 10 | 0 / 10 | 0 / 10 | 0 / 30 |
| Plus plus | 5 / 10 | 6 / 10 | 7 / 10 | 18 / 30 |
| While | 0 / 10 | 4 / 10 | 0 / 10 | 4 / 30 |
| Switch case end switch | 0 / 10 | 0 / 10 | 0 / 10 | 0 / 30 |
| Dot | 2 / 10 | 4 / 10 | 0 / 10 | 6 / 30 |

From the figure above, it can be noted that Google Cloud Speech API performs consistently better than the other two in most of the cases of keywords. The difference between Google Speech Recognition and Google Cloud Speech API lies within the fact that the latter allows for a list of preferred phrases to be passed into the recognition process, which greatly enhances the accuracy of the speech recognition. On the other hand, Microsoft Bing Voice Recognition seem to be only accurate for a few keywords and is largely inaccurate for most of the keywords involved.

### 3.2.2 Experiment with script for a sample program

The previous experiment focuses more on keywords, or very short phrases, as a starting point for our analysis. We realized that the actual recognition program will require the user to speak in complete sentences, for instance, to declare a variable, the user would say “declare integer count equal zero” instead of just “declare integer” as per our previous experiment. In order to affirm that Google Cloud Speech API is the most accurate API to use for our speech recognition, another experiment will be conducted. In this experiment, a script was written for a sample program to find the maximum number in an integer array, using my structured language (which will be presented in a later section in this paper). The same 10 subjects used in the previous experiment were tasked to read from the same script and their voices are recorded. Similarly, the audio files are then passed through the Speech Recognition module and the results are tabulated as follows in figure 3.

Note: The script consists of 64 words. The results below is the sum for 10 subjects, and the numerator represents the number of correctly recognized words for 10 subjects, while the denominator represents the total number of words for 10 subjects.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Google Speech Recognition** | **Google Cloud Speech API** | **Microsoft Bing Voice Recognition** | **Total** |
| Total | 389 / 640 | 423 / 640 | 247 / 640 | 1059 / 1920 |

*Figure 3: Summarized results of Speech Recognizer module with script for sample program*

The numbers above reaffirms that using Google Cloud Speech API produces the most accurate results over the other two APIs. Moreover, I have also observed from the raw data of the results that Microsoft Bing Voice Recognition is more suitable for recognizing proper complete English sentences than user-defined structured language. As such, I have decided on the use of Google Cloud Speech API for my speech recognizer module.

## 3.3 Structured Language

I have defined a set of structured language in order to express programming constructs in an unambiguous manner so that the computer program can accurately translate them into program code which is intended by the programmer. To enable programmers to remember and use the structured language in a fluent manner, some natural English constructs are also incorporated into the structured language as well. The current state of the structured language is defined below in figure 4.

|  |  |  |  |
| --- | --- | --- | --- |
| **Programming Constructs** | **Structured Language** | **Example speech** | **Example code** |
| Verbalizing words of different forms | Integer: just verbalize the number | Ten  One hundred and thirty four | 10  134 |
| Float: verbalize the number and use point for the decimal place | Thirty four point two | 34.2 |
| String: String (string content) end string | string hello end string | “hello” |
| Character: character (character content) | character x | ‘x’ |
| if-else conditions | begin if – then – else - end if  or  begin if – then – end if  (When dealing with nested if-loops, use end if to close the if loop) | begin if A then B begin if C then D else F end if G end if | if (A) {  B  if (C) {  D  } else {  F  }  G  } |
| Variable declaration | declare (variable type) (variable name) end declare  or  declare (variable type) (variable name) equal (literal) end declare | declare integer x end declare | int x; |
| declare float tax rate equal one point zero seven end declare | float taxRate = 1.07; |
| Array declaration | declare (variable type) array (variable name) size (number) end declare  (Using ‘with’ is optional) | declare integer array sequence size ten end declare  declare integer array sequence with size ten end declare | int sequence[10] |
| Variable assignment | (variable name) equal (literal) end equal | x equal ten end equal | x = 10 |
| Array assignment | (variable name) array index (number) equal (literal) end equal | sequence array index three equal four end equal | sequence[3] = 4 |
| Function declaration | create function (function name) return type (variable type)  0 or more: [parameter (variable type) (variable name)]  begin – end function  (Using ‘with’ is optional) | create function search return type integer parameter integer lower parameter integer higher begin A end function  create function search with return type integer with parameter integer lower with parameter integer higher begin A end function | int search(int lower, int higher) {  A  } |
| Function call | call function (function name)  0 or more: [parameter (variable name)]  end function  (Using ‘with’ is optional) | call function search parameter lower parameter higher end function  call function search with parameter lower with parameter higher end function | search(lower, higher) |
| For loops | for loop condition  (variable name) equal (literal)  condition  (variable name) (comparison operator) (literal)  condition  (variable name) (operator) begin – end for loop | for loop condition i equal one condition i less than sum condition i plus plus begin end for loop | for (i=1; i<sum; i++) {  } |
| While loops | While (variable name) begin – end while  While (variable name) (comparison operator) (variable / literal) begin – end while | while x begin B end while  while x less than two begin B end while | while (x) {  B  }  while (x < 2) {  B  } |

This set of structured language will be edited from time to time in order to improve speech recognition as well. For example, the Speech Recognition module has problems trying to decipher the word “if”. In most cases, the word “if” is omitted by the Speech Recognition module totally. The language is then changed for that particular construct, in the sense that the user is required to say “begin if” instead, as the Speech Recognition module has a higher accuracy in recognizing “begin if” as compared to “if” alone. More adjustments were also made during the implementation process, where signs of ambiguity were discovered in the structured language.

*Figure 4: Structured language for Coding by Dictation*

## 3.4 Word Corrector Module

Our Speech Recognition module is not perfect, we cannot 100% translate voice into text accurately. There will be instances where the Speech Recognition module misinterprets the audio. As a result, there is a need for us to process the input text into processed text, and this is where we utilize our Word Corrector module. There are 2 phases in this Word Corrector module. The first phase focuses on correcting the common errors which are misinterpreted by the Speech Recognition module, whereas the second phase focuses on using contextual information about previously declared variables to perform the correction of variable names used in the future part or lines of the coding process.

### 3.4.1 Word Corrector Phase One

In the first phase of the Word Corrector module, we correct the common errors which were misinterpreted by the Speech Recognition module. We utilize the raw data results retrieved from our earlier experiments on the basket of keywords and script for the sample program to perform our word correction. For instance, we realized that the end constructs are often misinterpreted (i.e. “end function” misinterpreted as “and function”), therefore we will attempt to correct “and” into “end” if the word “and” appended with the following word will form a valid end construct. As such, the Word Corrector module mainly targets the correction of program keywords only. Currently, the module uses a hardcoded algorithm to perform the correction of common errors, which will be improved upon in future. This will be discussed later in another section of the paper.

### 3.4.2 Word Corrector Phase Two

The second phase of the Word Corrector module uses contextual information about previously declared variables to perform the correction. After declaring variables, we will expect them to be used again, and variables which are not declared should not appear in the code at all as it will cause a syntactical error for the program. Thus, declared variable names will be stored in a list and subsequent use of variable names will reference this list to correct any erroneous variable names. This correction is done by finding the most similar word in the list to the word recognized by the Speech Recognizer module, and replaces that word with the most similar word found. The most similar word is found with the aid of the jellyfish python library, which does approximate and phonetic matching of strings (Turk and Stephens, 2016).

Firstly, we would convert all of the words into their phonetic encodings using American Soundex, and then compare the resulting phonetic encodings with Jaro-Winkler Distance to get a similarity index between two words. These are done with the help of the jellyfish python library. As the jellyfish python library provides more than just one way to doing phonetic encoding (namely: American Soundex, Metaphone, NYSIIS) and string comparisons (Levenshtein Distance, Damerau-Levenshtein Distance, Hamming Distance, Jaro Distance, Jaro-Winkler Distance), there is a need to test which pair of algorithms deliver better results. It is preferable to have higher similarity indices for closer matching words phonetically, and differentiable similarity indices between different pair of words (i.e. no ties in similarity index). After much testing, I have found that American Soundex and Jaro-Winkler Distance are the best fit for this purpose.

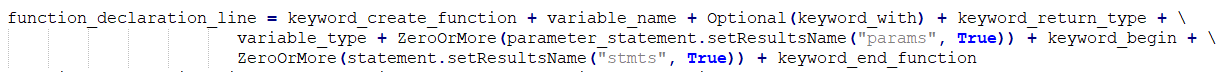
It is also worth to note that the way which we use to find the closest matching pair of words is actually a heuristic and is not an algorithm. This is due to the fact that it is difficult to truly tell which word is closer to a particular target word. For instance, if we have already declared the variables “eye”, “ai”, “aye” and we need to correct the word “i”, which of these 3 variables would we deem as closest matching to the word “i”. However, this heuristic is good enough for the purposes of this program in most instances, after testing the program with our script for the sample program from our earlier experiment.

## 3.5 Word Parser Module

This module translates processed text from the Word Corrector module into structured command for the Structural Command Parser. This module uses the pyparsing python module (McGuire, 2016) which allows parsing with user-defined grammar rules, in this case, our structured language, and converts it into structured command. More on structured command will be covered in the next module in the following section. The usage of a parsing library like pyparsing here is to ensure the robustness of the system and allows for easy editing of the structured language whenever necessary. This is a stark improvement from the initial implementation of the parser where parsing is done in a hardcoded manner.

We have also introduced the concept of context free grammar in defining our structured language’s grammar rules, after realizing that the initial set of language defined is not really properly designed and errors are not properly detected. With context free grammar, the grammar rules are now more well-defined and that new constructs or edits can be done more easily to the parser now.

The figure below shows an example of the function declaration construct using pyparsing and using context free grammar.



*Figure 5: Python code for function declaration construct*

From the figure, we can see that the structured language defining the function declaration construct can be easily changed and it is properly defined from the start keyword to the ending keyword involved. Any part of the construct can be replaced by some other keyword if we want to change the keyword to use for this construct, and we can also add in extra keywords or remove some terms in the construct should the need arises. Furthermore, context free grammar allows for recursive definitions, which allow us to create nested constructs like nested loops for example. In the case of nested loops, a for loop will allow for statements to be inserted into the for loop body, and a statement can be another for loop as well. All these increases the robustness and flexibility of the system.

## 3.6 Structured Command Parser Module

Structured command generated from the Word Parser module is converted into program code here. This module is basically a wrapper module which runs the Java program from the previous project ‘Talk-to-Code: From Structured Command to Source Code’ (Gao, 2016) as mentioned earlier. His program is able to construct a compilable and runnable C program with its Abstracted Syntactical structure, after accepting his self-defined structured command input, which we have converted from our structured language. This module completes our end-to-end program and generates C program code.

# 4 Plan for the next semester

For the next semester, I will continue to improve on the project to make it easier for programmers to code with their voices, so as to empower CTS patients to write code once again and also to reduce the risk of contracting CTS. The areas of improvements and their details will be elaborated on in the following subsections in this paper.

## 4.1 Improving the Structured Language

The current structured language might be a little bit hard to remember by heart and may still lack some natural English language aspects, in the sense that it is not really spoken in a natural English manner. Thus, the plan for next semester would be to continue redefining and editing the structured language, in order to make the speaking more natural, while retaining the unambiguous domain of the language so that the translation to program code is not ambiguous. There will also be attempts to improve the structured language to use keywords which can be more easily recognized by the Speech Recognition module to make the recognition process more accurate.

## 4.2 Adding more constructs to the Structured Language

The current structured language does not contain every programming construct available, and is restricted to only a limited set of constructs. The plan for the following semester would be to introduce more constructs to the program. Some of the constructs which were planned to be added are as follows:

1. Switch – Case construct (The switch statement)
2. The use of logical “and” and logical “or” when performing comparisons
3. Shorthand assignment operators (such as “+=”, “-=”, “\*=”, “/=”)
4. Allowing the use of symbols like percentage (%), dollar ($) and ampersand (&)

The sets of constructs to be added are not exhaustive. There will be many more other constructs which could potentially be added as well, and the more commonly used constructs would take precedence in the order of my implementation in future.

## 4.3 Improving Word Corrector Module

As mentioned earlier, the Word Corrector module’s phase one currently uses a hardcoded algorithm to correct common errors, and we will improve on this aspect in the next semester. A hardcoded algorithm is not robust enough and is not easy for us to edit or extend the current state, as such a better algorithm is necessary to improve our module further. One possible way of doing so would be to make use of the word similarity heuristic, as presented in Word Corrector module’s phase two, to execute the correction. We could possibly use a list of keywords and check if the recognized word from the Speech Recognition module has a close similarity to any of the keywords in the list and perform the correction. In the next semester, I will look at the plausibility of employing such a method to perform the correction and also investigate and research for more alternative methods to improve the correction, to check if there is any better way of doing the correction.

## 4.4 Improving overall program and feedback

The current Coding by Dictation application is a simple program which converts voice input into program code, and it has very minimal User Interface (UI). These aspects would be worked upon in the next semester to give the user a better coding interface. Moreover, there is only a limited set of feedbacks given to the user according to what the user has input into the program. There is a need to give proper and more detailed feedback to the users, which will greatly enhance the coding efficiency of the programmer. If time permits, I will also look into allowing the user to edit the program as that is more similar to the thought process of a programmer. A programmer does not code by writing out a whole piece of code on paper first before typing it into the computer. The thought process would include going back in the middle of coding to declare a variable which is needed to be used at the current stage of coding, thus the need for user to add or edit code previously entered. This will create a more complete program which would resemble closely to how programmers would usually type code.

# 5 References

Chew, Y.X. (2016). Hands Free Programming. National University of Singapore, 2016

Crouch, Tammy (1995). Carpal Tunnel Syndrome and Repetitive Stress Injuries: The Comprehensive Guide to Prevention, Treatment and Recovery. Frog, Ltd. Berkeley, California.

Gao, R.S. (2016). Talk-to-Code: From Structured Command to Source Code. National University of Singapore, 2016

McGuire, P. (2016). Pyparsing (Version 2.1.9) [Software]. Available from <http://pyparsing.wikispaces.com/Download+and+Installation>

Turk, J. and Stephens, M. (2016) Jellyfish (Version 0.5.6) [Software]. Available from <https://github.com/jamesturk/jellyfish>

Zhang, A. (2017). Speech Recognition (Version 3.7) [Software]. Available from <https://github.com/Uberi/speech_recognition#readme>

1. Read http://advice.writing.utoronto.ca/types-of-writing/literature-review/ to find out

what literature reviews should and should not be. You may want to check with your

supervisor on the best way to do a literature review in your research area.

2. Make sure you are aware of what constitutes plagiarism. See

http://www.cdtl.nus.edu.sg/success/sl7.htm.

This is a gentle reminder for you that the deadline to submit your CA reports directly to your respective supervisor and main evaluator is **5.00pm on 8th November 2017. (10 – 15 pages)**