MATLAB Speech Synthesizer

Final Report

Lachlan Dow

2021

**Abstract**

This paper will describe design, development and testing of a formant speech synthesizer built in MATLAB. The process took place between the 28th of October 2020 and the 27th of April 2021. The design process took place from the 28th of October to the 19 of February 2021 the implementation took place from the 20th of October to the 30th of March and the testing took place between the 2nd of April and the 24th of April.

**Overview of the System -**

**Background and History of Speech Production Systems-**

For hundreds of years, people have been trying to mechanically produce human speech. Early models created in 1779 by Christian Kratzenstein (History and Development of Speech Synthesis, 2021) generated vowels by constructing models that resembled the human vocal tract (as mentioned above) by vibrating reeds like musical instruments.

Diagram

Description automatically generated

Figure 1. Katzenstein’s resonators (History and Development of Speech Synthesis, 2021)

In the subsequent century and up until the 1960s, multiple models for mechanical and semi-electrical analogues of vocal systems were created. The most notable of these systems is known as the VODER (Voice Operating Demonstrator), which was developed by Bell Laboratories in the mid-thirties (History and Development of Speech Synthesis, 2021). The device was designed to analyse speech into slowly varying acoustic parameters that could then drive a synthesizer to approximately reconstruct the original speech signal.

Diagram, schematic

Description automatically generated

Figure 2. The VODER speech synthesiser (Klatt, 1987)

The invention and demonstration of this system caused the scientific community to become increasingly interested in speech synthesis. In the coming decades, systems like the Pattern Playback synthesizer, constructed by Franklin Cooper (Klatt 1987), reconstructed recorded spectrogram patters into sounds. The first formant synthesizer, PAT (Parametric Artificial Talker), was constructed by Walter Lawrence in 1952. This system consisted of three electronic formant resonators connected in parallel. Moving glass slides where then used to convert painted patterns into six-time functions to control different parameters for the synthesiser. Around the same time, a synthesiser that processed sound using cascade resonators was introduced called OVE I (Orator Verbis Electris). This system would go on to be redesigned and improved over the text two decades in the OVE II and OVE III.

These two systems were combined by John Holmes in 1972 when he introduced his own parallel formant synthesiser. He was able to tune the synthesiser to say “I enjoy the simple life” well enough that most people couldn’t tell between the synthesiser and a real voice.

In 1979, the MITalk system was developed by Allen, Hunnicutt and Klatt. This system used all the advancements up to this point including the parallel/cascade synthesis method to produce extremely realistic speech sounds. This system would later go on to be adapted to be used in DECtalk – a commercial text to speech device.

Text-to-speech synthesis has continued to advance to the stage it is at today, with complex unit synthesis systems that use POLSA technology to blend pre-recorded text together in an extremely realistic way. These systems are getting so to the stage where it can sometimes be difficult to differentiate between them and a real human talking.

//TODO: add more details about modern synthesis technology

**Design –**

The design process began with investigation into speech synthesizers, their history, and current state of the art synthesizers. This stage uncovered multiple different types of synthesis, namely:

* Unit selections synthesis, which uses a large database of recorded speech when each recorded utterance is segmented in different ways using a specially designed speech recognition system. These segments are indexed based off defining factors e.g., fundamental frequency, duration etc.
* Diphone synthesis, which uses a database containing all the diphones occurring in a language. Only one example of each diphone is contained in the speech database, then at runtime the prosody of a sentence is superimposed on the units, usually through a linear predictive coder (LPC).
* Domain-specific synthesis, which concatenates pre-recorded words and phrases to create complete utterances, is generally only used in speech variance as it will be very limited, for example a train announcement system or weather reports.
* Formant synthesis, this method uses an electronic speech production model where different parameters are varied to create speech sounds, these values can then be interpolated to create diphthongs or stopped to create plosive bursts. This method is said to create a sound that is very easily understood but sounds very robotic.

Until recently, this was the best option for synthesis due to its customisability and adaptability.

In the initial stages of design, creation of a unit selection synthesis system was investigated as a strong possible solution. In the experimentation stage, MATLAB was used to take speech segments and separate them into their respective sounds, and then reassemble them into the sounds that were needed to create other words.   
After more investigation into other options and reading speech synthesis and recognition (John and Wendy Holmes) and MITalk, this method had more information on the use of formant synthesis as a speech production method, and coupled with advice from the project advisor, a decision was made to transfer the project to a formant synthesizer.

Predesign for the MATLAB program was split into two parts-

1. The formant synthesizer, based off the parallel/cascade design(Klatt. D, 1980)
2. The text pre-processor for input into the synthesizer

**The Formant Synthesizer – Design and Development**

The design of the formant synthesizer was adapted from the design found in the paper by Klatt himself and also other adapted versions that were more suited for modern systems, most notably the system developed by d'Heureuse (2019). The synthesizer was **Diagram

Description automatically generated**broken down into different sections as objects in MATLAB, and each object would have step functions to replicate a signal passing through them and the effect that it would have on that signal.

Figure 3. Synthesizer class diagram

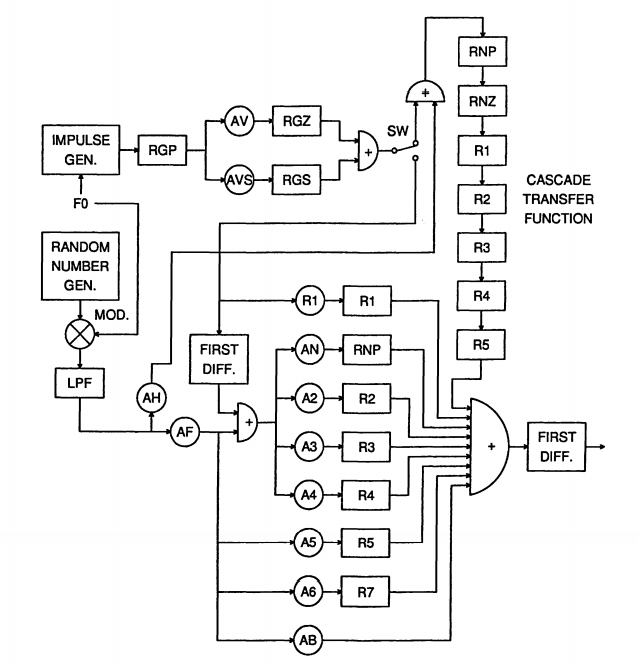


Figure 4. Block diagram of Klatt cascade/parallel formant synthesizer.

Figure 2 shows the model of the Klatt formant synthesizer proposed in his paper (Klatt, D., 1980.)

Pictured Above is the class diagram for the synthesizer in MATLAB. This has been built as an object-oriented application. The main function for this system is the generator object, which is responsible for creating the sound waveforms from the parameters inputted. The main function is the generateSound() function. The function takes a data frame as input as well as the main parameters for the synthesizer, and then will process the values for the sound through each branch to output an array of doubles that can be used to produce a sound using the sound values.

**Difficulties in creation –**

When constructing the synthesizer, a waterfall method of production was used. Generally, to have a system that produced any sound at all, the whole base structure of the synthesizer would be needed. Because of this method, after laying out the base structure of the synthesizer roughly 4 weeks was spent debugging; stepping through the program and ensuring variables were correct at each stage as well as fixing errors that were caused by putting all the different classes together. In the future, it would be good to construct a predefined testing plan for each function and class to ensure that they would work correctly with the synthesizer. This would positively impact the production of the generator class, as that was where a large portion of the errors were found.

Generating white noise - this problem was discovered as the random number generation in MATLAB has a very wide spread of differences that tend to gravitate towards the more extreme ends of the scale that has been given. This caused a lot of confusion in the initial development of vowel sounds. Multiple options were tried including another random number generation method but in the end, simply reducing the variance in the numbers was enough to make the sound understandable.

Nasalisation – when attempting to create nasalisation in the synthesizer, a buzzing sound is produced that made the important section of the sound inaudible. This was fixed by adjusting the nasal antiformant resonator to remove a larger section of the non-nasal sound. This fix came with the effect of the sound’s quality seeming slightly hollow but it increased legibility to a large degree.

Pseudocode for Text Reformatting:

1. findAbbreviations()
2. split array into words
3. loop for each word
   1. use regular expression to find acronyms
   2. regular expression to check for numbers
   3. if the word is a acronym
      1. replace the acronym in the array
   4. end if
   5. if the word is a number
      1. replace the number with the word equivalent
   6. end if
4. end for loop
5. change all words into the lowercase versions for the next stage of pre-processing

**Diagram

Description automatically generatedPre-Processing unit**

The pre-processing unit was developed using an agile approach. Each part was developed and functional in a deployable way. This section of the system was broken down as:

1. Text Reformatting
2. Text-to-phoneme conversion
3. Phoneme-to-sound frame conversion
4. Sound duration computation

More detail on these sections is provided below.

**Text Reformatting**

In this section, numbers were broken down into their word equivalent with a series of functions and acronyms were translated into individual letter segments to get the right sound from the system, and some abbreviations were translated into their full words.

**Text-to-Phoneme Conversion**

Pseudocode for Text-to-Phonene conversion

1. while loop for every letter
   1. get index position of rule in array
   2. loop until match found
      1. extract the character matches for rule segment
      2. check if the segment matches the section of text from current letter
      3. if this section matches extract the conditional rules for match before letter rules
      4. check that the string preceding the rule set matches the extracted rules
      5. extract conditional rules after the matched letter set
      6. check the string after the rule set matches the extracted rules
      7. assign Boolean value to matchFound
      8. if match is found push phonemes equivalent from the rule to array of phonemes
      9. increase rule index value by one and return to start of array
   3. end loop
2. end while

This function’s purpose was to break down a word into its different phoneme segments. In an overarching method, the function word2Phone will take a word input, a map of key-value pairs for the different letters, and their position in an array of letter-to-phoneme rules. The function then loops through every letter of the word, referencing the word conversion with the rule list, and creates a new array with the phoneme conversion.

Pseudocode for Phoneme-to-Sound conversion

1. loop for the length of the array of phonemes
2. if sound doesn’t contain diphthong
   1. if sound is a \_ for HH sound
      1. get next phoneme and convert to HH equivalent
      2. add sound to sound array
      3. move onto phoneme after adjusted one
   2. end else
      1. convert to sound
      2. move onto next phoneme
   3. end if
3. if sound does contain a diphthong
   1. split diphthong into segments
   2. convert phoneme into sound frames
   3. assign diphthong attribute to these sounds
   4. more onto next sound in the array
   5. end if
4. end loop

**Phoneme-to-Sound Frame Conversion**

This function will convert the phonemes into their sound frame equivalents. The function will loop through the array of phonemes and then convert them, using a map of the phoneme symbol, to an index representing the sound frame for that phoneme.

Pseudocode for Sound Duration Computation

1. check if array length is one
2. if the array length is one
   1. assume that the item is a letter and adjust the interpolation accordingly
   2. function return
3. end if
4. loop for length of sound array
5. if sound is last in array or plosive
   1. generate sound frames with no blending
   2. else
   3. generate frames for current sound frame generate blended sound for next frame
6. end if
   1. end if
7. end loop

**Sound Duration Computation**

This function is used to calculate the duration of sounds and how to blend them. The parameters inputs for this function are a list of sound objects to be converted. The function will then process these sounds into 5ms sound frames from durations based off a map of phonemes and duration characteristics then output these sounds as an array to be passed into the generateSound() function in the Generator class.

After these functions are run, the text is ready to be sent to the synthesizer.

Pictured in figure 3, is the diagram that was used to construct these functions and how thy fit together to be input into the synthesizer.

**Difficulties in Implementation of the Pre-processing system -**

One of the notable roadblocks encountered while implementing the pre-processor section of the synthesiser was interpolating between the different frames for the different phonemes. The aim was to interpolate between these data frames in a way that made sense and didn’t make the system sound like it was just jumping between the values too quickly. Yet, the transition could not be drawn out for too long or it would sound forced. This meant that values had to shift on a curve. To do this, weights were applied to the specific values and then the values were interpolated between each of these including the weights, so as the value got closer to either end, the interpolated values would get nearer to the end values faster. //TODO: rewrite better

Another notable difficulty was in translating a number into words. The function for this was lengthy and complex, and required many switch statements to handle the length of the number. The way this was implemented in the end was with switch and if statements, the pseudocode below shows this process:

The main difficulty in this code was creating the conversion from a three-digit number to their word equivalent. This code was very large and contained a lot of switch statements with cases for each number in each level of the number e.g. [one hundred and], [twenty], [one]. Looking back, it would have been far simpler to break the number down into its respective components then, using maps, to test the components and simply insert strings accordingly. Maps could also be reused for the hundreds and unit sections of code.

Pseudocode for conversion from number to sound

1. if number contains a decimal point, split at decimal
2. find modulus 3 of number length
3. loop over number
   1. if there is a remainder
      1. extract the section of the start of the number
      2. convert the number into the word equivalent
      3. test the remaining length of the string to determine what denomination should be added to the other strings(million ect)
   2. move across the string for the length of the remainder
   3. else loop for the other section of string
      1. extract the section of the start of the number
      2. convert the number into the word equivalent
      3. test the remaining length of the string to determine what denomination should be added to the other strings(million ect)
   4. move across the string for length 3
   5. end if
   6. if there is a decimal point
   7. add the word point
   8. loop for every number after decimal point
   9. get number equivalent for the number
4. end if
5. join all the word into one string
6. split number into its different digits
7. loop for length
   1. if length == 2 and number >=10 AND < 20
      1. map digit to teens digit
      2. return
   2. end if
   3. if remaingLength 3
      1. get value from map of digit
      2. add “hundred and” to the text
   4. else if remainingLength = 2
      1. get value map of tens digit eg “thirty”
   5. else if Remaininglength == 1
      1. get value from map digit
   6. end if
8. return combined number string

**Functionality testing of the system -**

In order to test the different sections of the system, using test cases for the different methods and classes was required. Testing the pre-processor section consisted of running individual methods with expected input

**Conclusion -**

Working Klatt formant synthesizer – 12/2/21

Text Pre-processor – 26/2/2021

Creation of the user interface linking of Text pre-processor and formant synthesizer– 5/3/2021.

Testing and refinements – 15/4/2021

Report finalisation- 27/4/2021

With this schedule in mind the creation and completion of the synthesizer should be completed on time with all the main features mentioned before implemented.

**References**

Cs.mcgill.ca. 2021. *Speech Synthesis*. [online] Available at: <https://www.cs.mcgill.ca/~rwest/wikispeedia/wpcd/wp/s/Speech\_synthesis.htm#:~:text=The%20first%20computer%2Dbased%20speech,the%20history%20of%20Bell%20Labs.> [Accessed 26 January 2021].

Klatt, D., 1980. *Software For A Cascade/Parallel Formant Synthesizer*. [online] Fon.hum.uva.nl. Available at: <https://www.fon.hum.uva.nl/david/ma\_ssp/doc/Klatt-1980-JAS000971.pdf> [Accessed 20 January 2021].

Wiki.inf.ed.ac.uk. 2021. [online] Available at: <https://wiki.inf.ed.ac.uk/twiki/pub/CSTR/Speak14To15/evaluation.pdf> [Accessed 27 January 2021].

Jonathan Allen, M. Sharon Hunnicutt, Dennis H. Klatt, Robert C. Armstrong, and David B. Pisoni. 1987. *From text to speech: the MITalk system*. Cambridge University Press, USA.

d'Heureuse, Christian. "Chdh/Klatt-Syn". *Github*, 2019, [https://github.com/chdh/klatt-syn. Accessed 4 Feb 2021](https://github.com/chdh/klatt-syn.%20Accessed%204%20Feb%202021).

Research.spa.aalto.fi. 2021. *History and Development of Speech Synthesis*. [online] Available at: <http://research.spa.aalto.fi/publications/theses/lemmetty\_mst/chap2.html> [Accessed 12 April 2021].

Klatt D. (1987) Review of Text-to-Speech Conversion for English.*Journal of the Acoustical Society of America, JASA* vol. 82 (3), pp.737-793.

|  |
| --- |
| . |
|  |