Speech Synthesis of the Modern Age

Adam Stammer

Table of Contents

1.1 – Introduction

Speech has long been one of the primary tools of society. Any successful grouping organism has some way to communicate information, from the honey bee's waggle dance to the whistling of dolphins. Humans developed speech for this purpose. It passes more information than body language alone and often does so much faster and at greater distances. The desire to (re)produce speech by means other than natural human interaction is obvious. Today machines everywhere speak to users. Smartphones, GPS machines, automated phone systems, cars, computers, and many other systems read things off to people on a daily basis.

2.1 – The Process

There are many different ways to produce sound that resembles speech, but they all follow the same general process. The depth and skill by which these steps are carried out are what make one speech synthesizer more appealing or accurate than another. Since the beginning of speech synthesis the process itself has changed, and been strayed from very little, but the methods of execution have gotten increasingly more complex and precise.

2.2 – Text Interpretation

By many this is seen as the easiest of the steps. If given text, tell me what it represents; what are the words? For a human this is generally very easy, but for a machine this can be far harder than it seems. At the start it is quite easy because words are generally separated by spaces. Most systems would then lookup the given groupings in a stored dictionary of words to verify not only that the word exists, but also that it only has one meaning.

One of the problems faced is homographs, words that have more than one pronunciation depending on their meaning, "read" for example. Context is how humans know the proper way to say

such a word and so part of the purpose of this step is to build the context. This is when it is decided what tense the words are likely in.

Another purpose of this step is to quantify characters, numbers, and special characters. Given the number 321 it could represent many different things. It could be a simple quantity: “three hundred and twenty one”, but it could be the year “three twenty one”. Or maybe it’s a pin or part of a countdown: “three two one”. Again, context is the key. Words indicating time like “year”, “born”, “died”, “in”, “from”, etc. seen in the same sentence likely mean it should be pronounced as a year. Likewise, words indicating quantity like a following plural noun certainly imply a different pronunciation. Punctuation and special characters like “$”, “@”, “%”, “&”, “( )”, etc. also have their own forms of ambiguation. Currency symbols can certainly help give context to numbers but how does one verbally indicate things in parentheses, like example lists. The asterisk is often used many different ways, sometimes as multiplication, other times in way similar to parentheses. It is often difficult to express characters like these that more often than not only express things by proxy, through affected words and phrases.

Simply reading the words on a page may seem elementary to the veteran reader, but it is no small task. Problems can crop up anywhere; a single character can completely change the meaning of a sentence. But it this first step that aims to work through that. Work through the text to disambiguate all of the nuances of writing. Use the pieces to build context to better understand what all of the pieces mean individually.

Some speech synthesis systems do not automate this step and instead rely only on ‘hardwired’ input. These would be systems not of text-to-speech but systems simply capable of making speech, either a set output limited to only a set few words or phrases, or something manual in input likely not computer based like using instruments to produce speech [1].

2.3 – Word Interpretation

Once the words have been interpreted and disambiguated the next step is to decide how to say them. This done by breaking the words into their given phonemes, the base units of pronunciation. This is when things not properly disambiguated in the first step will become a problem.

This step is often done with a dictionary. Looking up a given word will give you the phonemes that make it up. This approach is very simple and assures that the words being said are real and are pronounced correctly, but it limits the possibilities of the synthesizer when it comes to new, unrecognized, or foreign words.

A common solution to this problem is to instead break words up into their graphemes similar to phonemes but written and interpolating the phonemes from them. This gives a synthesizer the ability to pronounce new words, often with reasonable accuracy. Proper nouns often see the most benefit from this design as many names and companies do not make it into word dictionaries. It does, however, leave the system more vulnerable to mispronunciations. Some English words like “island”, “coup”, or “yacht” would likely be mispronounced. Another benefit of this method, though, is that it doesn’t require a dictionary of words which can be quite valuable when storage space is at a premium. Because of this they are often considerably faster as well.

A hybrid system is often seen as the best option. Using a word dictionary ensures that most words are pronounced correctly, even the oddly spelled ones, but using a grapheme system as a fallback for when a word is not present in the dictionary means that new or unrecognized words won’t be completely skipped over. This also gives the option for removing some words from the dictionary so that storage space isn’t entirely compromised [1].

2.4 – Phoneme Interpretation

This final step often has the strongest effect on the “realness” of the final output. It decides not only what the voice will sound like (male/female, deep/high, rough/smooth, etc) but also how smooth the sounds come out. Speech synthesis is often criticized as too robotic and emotionless. Since it is this step that actually produces the sound, it is most often this step’s fault for an unrealistic voice, or this step’s praise for one that is. Turning phonemes to sound has also been one of the more varying aspects of this process as developing hardware and software often have the strongest impact at this point. Because of this all of the methods can’t be covered in this paper but the three most common will be [1].

2.5 – Articulatory Production

By far the most difficult and complex of methods articulatory synthesis. It is any means of phoneme production by physical mechanical means. One of the first voice synthesizers created used a system of gears, pulleys, and bagpipe-like wind pipes to produce sound that was apparently recognizable as speech. Modern approaches are far more advanced using designs based off of the human vocal system, using various electrical components to act just like our throat, mouth, and lungs work to make speech. While this should produce the most realistic speech it remains the least researched and experimented because of its complexity and difficulty of adaptation. A replicated voicebox cannot be shrunk to fit inside of a smartphone and so the practical uses remain confined mostly to humanlike robots, a field still lacking in development [3].

2.6 – Concatenative Production

By far the simplest method, and the one in most use, is to use prerecorded phonemes and simply concatenate them, or piece them together in order. This allows a recording from one person of the roughly forty phonemes used in the English language and, using a speech synthesis system pronounce say any word.

Due to the piece by piece nature of this method, the end result is often choppy if not done well and because phonemes often sound different based on their surrounding phonemes the voice produced is usually not expressing the proper tone, if any at all. Some more advanced concatenating systems will record multiple phonemes more than once and use the appropriately toned phonemes to add a little more emotion to the voice. Some systems also use multiple recordings of phonemes at random to add variety within the generated speech. This often adds more realism to the end result at the cost of storage space, and recording time [3].

2.7 – Formant Production

The last common method of producing sound from a phoneme involves an analysis of the sound itself. Rather than using recordings to build sound, digital hardware generates the audio by using the proper frequencies and combines them to produce the phonemes. This requires much more powerful hardware than concatenative systems do and there is a substantial market for processing chips designed specifically for voice synthesis.

One of the major benefits to this method is that you are no longer limited to the prerecorded phonemes. This can make foreign words more likely to be pronounced correctly. GPS devices and translators are often equipped with Formant based synthesizer because of this. Because all of the sound is generated dynamically the need for storage space is also reduced drastically. On the downside, this method generally produces much more robotic voices that other methods because there is no actual human voice involved in it’s production [3].

3.1 – Modern Advances

In more recent times digital technologies have advanced considerably since the dawn of voice synthesis and the effects can certainly be felt. Stronger hardware has made Formant designs more and

more efficient and realistic, lab produced articulatory designs are more and more lifelike than ever before, and machine learning is being applied to create entirely new methods [3].

In 2016 DeepMind released a paper detailing their new neural network design, WaveNet, as applied to audio production. This method involves manipulating the raw audio, similar to the Formant method, but uses a trained neural network to decide the next audio piece. Each sample of audio is generated based on previous output of the same sequence, as shown below in Fig 1. It is much more demanding than even the most complex Formant systems but the results have shown voices that are far more realistic than previous known methods. It even mimics breathing sounds that most previous models don't. Because this method is working sample by sample through the raw audio, the preprocessing steps are much different than that of the methods discussed previously [2].

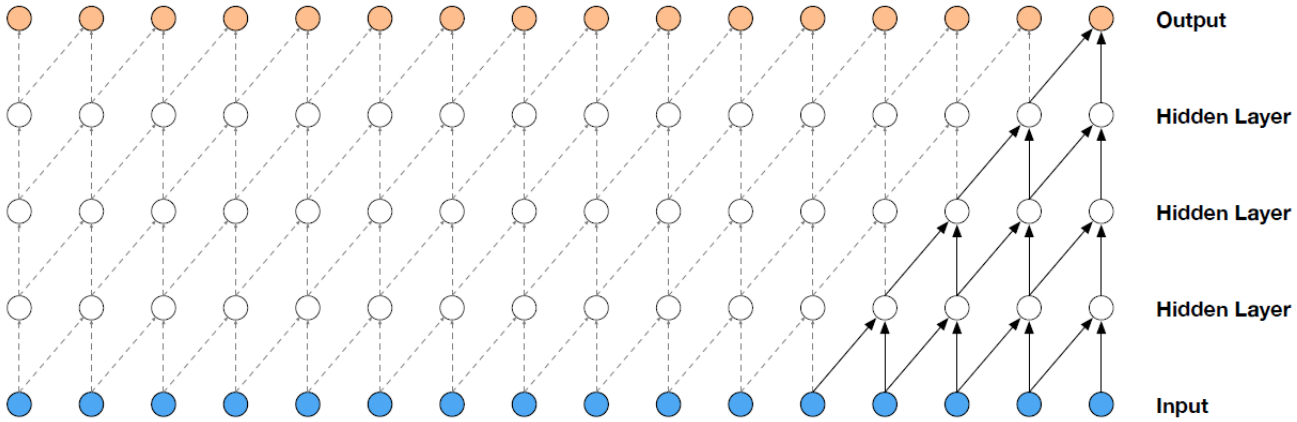

Fig 1 [2]

4.1 – Application and the Future

Although it has been referenced numerously already the applications of synthesized voice is extensive. Computers and smartphones are often equipped with a text to speech narrator for those that have trouble reading or seeing. GPS devices and vehicle computers come with hands free interaction systems to assist people with minimal interruption. Digital Assistants are appearing in more and more places, from phones, to home automation [3]. Voice synthesis has long been seen throughout science

fiction as sign of an advanced people. From 2001: A Space Odyssey's Hal to Iron Man's Jarvis the demand for generated speech is high and as such uses for it are being developed and produced almost as fast as the technology itself is progressing. Projections for technological development are universally expecting advances to continue, perhaps even speed up. This has the potential to push voice synthesis into a whole new level of realness, efficiency, and effectiveness.

5.1 – References

[1] C. Woodford, “How speech synthesis works,” *Explain that Stuff*, 08-Jan-2018.

[2] S. Sander, *WAVE NET: A GENERATIVE MODEL FOR RAW AUDIO*. London, UK: Deepmind Limited Technologies, 2016.

[3] T. Dutoit, “A Short Introduction to Text-to-Speech Synthesis,” *An Introduction to text-to-speech synthesis*, 17-Dec-1999. [Online].