Vocal Synthetics

Designing for an Adaptable Singing Synthesizer

by

Liam Clarke

A thesis presented to OCAD University in partial fulfillment of the requirements for the degree of Master of Design in Digital Futures

Toronto, Ontario, Canada, 2021

Abstract

Technological music tools such as digital audio workstations and electronic music instruments have enabled musicians without formal training to create music that is heard by millions of people. The automation by software and hardware can create compelling productions without limitations from performance ability. However, the automation of vocals is particularly difficult because beyond pitch and timbre, the vocalization of language requires additional parameters for control. As the production of a vocal synthesizer and its vocal palettes is complex, the current market sees these difficulties represented through products that have limited voices and do not adapt to vocal trends. This project demonstrates a tool that allows producers to use a simple typing interface for the input of words, allowing the output to be integrated and controlled by modern digital audio workstations. Using a machine learning solution, the tool is not dependent on large stores of audio data once a model is trained and since it contains a simple method to create new voices, it can keep up with evolving musical trends and vocal styles. The aim is to bring the human voice into the realm of digital music production enabling a music maker to include a large range of vocal styles within their production tool set. This paper outlines the design and development of the tool and culminates in a piece of music that illustrates the value of applying design thinking research strategies to an artistic and technical challenge.

Acknowledgments

I would like to express gratitude to the following:

Adam Tindale, for his expert and thoughtful guidance, through classes and late night conversations over the past two years.

Suzanne Stein for her valuable advice and patient feedback throughout this project.

Maya Malkin, for her talent.

Kate Hartman and the Digital Futures program.

My parents for their support.

This research was supported by an OCAD U Graduate Student Project Grant.

Table of Contents

Table of Figures

Introduction & Research Motivation

Introduction

My objective for this thesis project was to build an effective singing voice synthesizer that could be used by music makers to create vocal music and would feature a simple process for adding new vocal styles. My overall research question for this project was as follows: How might we create a convincing system that has value to a music maker using current speech synthesis technology? In this paper, I describe the inspiration for and how I built my singing voice synthesis music production tool through available open source resources surrounding speech and audio synthesis. The aim is to bring the human voice into the realm of digital music production in a framework that would enable a music maker to include a large range of vocal styles within their production tool set. Having a comprehensive and expanding digital bank of vocal styles would enable new creative possibilities without the complexities associated with traditional vocal production. The synthetic singing voice is not intended to replace a live performer. Rather it provides music makers without access to a specific singing ability or vocal style, the option and flexibility for inclusion within their work.

The developments in the field of machine learning open up the possibilities of creating a highly realistic vocal synthesizer. Machine learning is complex, especially within the context of audio processing and synthesis, for a designer who has not undergone formal training in the field. Therefore, the strategies presented in this work are not directly aimed at furthering technological advancements in neural systems for audio synthesis, but focus on applying creative design strategies within the existing advancements of speech synthesis to create a singing voice system.

As a music maker, I have used digital audio tools extensively. Using first Garageband and now Ableton Live, I compose and produce instrumental music for sale and performance through music libraries, record labels and collaborations with songwriters and singers. The digital tools available provide enormous flexibility to create a wide

range of different types of music genres. No longer is the music maker restricted to working with live musicians to make music. Many instruments now live in digital space where the music maker can be both the composer and player. There is almost no limit when it comes to the instrumental aspects of song crafting. Unfortunately, the almost limitless possibilities of computer-based music do not extend to the vocal portion of music. When writing songs, as a non-singer myself, I am dependent on collaborations with a vocalist to complete non-instrumental music. This can add enormous complexity and cost to the production and even the legal aspects (such as royalty splits) of the end-to-end process. Moreover, when working with an artist, I would like to be able to continuously experiment with the vocal melody myself, until I feel it's satisfactorily catchy. Throughout my music making career, I have looked for digital solutions that I could use as a substitute singing voice; solutions that are similar in flexibility to the tools available for a wide range of instrumentation and effects.

The advancements in computer power and, importantly, memory capacity, have led to powerful instruments built from sampling the recordings of live instruments. Getting the sound of a specific instrument or full orchestra within a composition does not require training in recording techniques, recording tools or studio expenses and, most importantly, using the sound of an instrument within a composition does not require formal musical training. However, the reproduction of non-digital musical instruments in a digital workspace requires large stores of audio. Each note and its many articulations must be recorded and compiled into a database to be called upon through an interface. For example, a professional grade string section sampler that I use, contains thirty-two articulations for each note. Consider the variables for a human singing voice. Adding to variations in frequency and articulation, there is also the variable of spoken language. A singing music production tool based on a library of audio samples where a user could have lyrical and pitch control would require an unrealistic amount of data and storage space. Technical solutions that have been developed so far use multiple tricks to limit the size of the audio base. These solutions often result in qualities that are unnatural to the human voice. A machine learning based solution has the potential to enable the development of a singing voice music tool that is not reliant on large stores of audio

once a model is trained. Further, if the system contains a simple method to create new voices, it can keep up with evolving musical trends and vocal styles.

The paper begins with a discussion of the motivation behind my research project. It then provides an overview of speech and singing synthesis technologies to give some background technology context. I then describe my research including the research questions, research through design with iterative prototyping and the results. I conclude with a discussion of my thoughts on the research and the project results.

Research Motivation

I am a music maker whose music career began very traditionally. As a child, I studied a number of instruments through the Royal Conservatory of Music as well as taking part in the music programs of my various schools. I would often experiment with music composition and production, using primarily the guitar and basic home recording technologies. After the purchase of a 2004 iMac, opening books of standard sheet music was a rare occurrence outside of studies. With the iMac came Garageband, a simple DAW (Digital Audio Workstation) that featured MIDI (Musical Instrument Digital Interface) sequencing, audio recording and basic audio effects. This was a paradigm shift for me in terms of my place in the field of music. Instead of being a single instrument voice among many, Garageband enabled me to be the whole orchestra playing self-composed music. Though Garageband was a limited introductory DAW, it felt like the potential for creation was limitless. Over the following years, I moved on to a more professional production suite called Ableton Live, and further away from playing any real instruments. I found that as my sound design and audio programming abilities improved, digital versions of guitars that used MIDI notation input could sound better than my own guitar recording as I wasn't limited by my playing ability. To make an instrument such as the guitar sound real, I program details that give a recording life, such as pick scrapes, fretboard squeaks, automating the pitch bends and running it through digital guitar amps. This process takes quite a bit of effort and time at first, but

once this system is built it can be reused, easily changing MIDI notes depending on the context. My computer music skills had surpassed my musicianship.

Over time I developed production skills into a career built around creating music for advertising and visual content, personal artist projects and production for other artists. When considering the crafting of modern vocal music, a key area of the work involves an aspect of the song known as the topline. The topline of a song is the vocal component over an instrumental bed, designed to catch the ear of the listener. A popular assembly line method for the top producers is to create a beat and send it out to a large pool of vocal talent, who try to work out the catchiest topline. This diversification increases the chance of finding a hit topline. Lyrics are sometimes written after a stream of melodic gibberish has been arranged. This is referred to as the track-and-hook method by John Seabrook, whose book, The Song Machine (2016), took a look at this assembly line format involving the division of labour across the creation of a record. While the track-and-hook method is available to sought-after producers with access to talent in large numbers, the assembly line also goes the other way. Producers, often up and coming, are presented with toplines that mostly contain complete lyrics over simple beats, where they compete amongst a talent pool to turn it into a hit song. Working with predefined toplines is primarily the way I've been a part of large projects.

Besides being in the assembly line, competing to manufacture parts, I also work with artists directly. This gives me an opportunity to mold the topline, try different melodies and rhythms with an artist's vocals. We can record over a track, I can analyze the melody, then give suggestions during a re-record. The vocals are malleable, but so is the instrumentation, in an evolving process that still retains aspects of specialization like track and hook. To experiment with the first take of the vocals, I push and pull the rhythm and pitch of the audio. Analyzing the original sung melody, changing the pitch of parts and readjusting the rhythmic qualities of a vocal take help me find that ear catching melody in an analytical and experimental way. After some time experimenting, I hand back a template for the singer to re-record their lyrics, a sing-by-the-numbers approach.

D.A.N.C.E., by French electronic act Justice, is an early example of a voice in a song that I wanted available for use within my own productions. The vocals are provided by London's Foundation for Young Musicians children's choir. Over the choir backing vocals, one young member is featured, who brings what sounds like a young Jackson Five era Michael Jackson to the record. It wasn't just the lead vocal that captured my interest, but I became set on incorporating a children's choir within my music as well. Both these would be difficult to source, finding a sound-a-like to one of the most iconic voices would be luck, and while easier to find a children's choir, there are logistical and financial problems to address when hiring and organizing the recording of a full choir. Time and financial investments are not unique to choirs, all vocal production faces similar challenges. The writing and recording of vocals is a practice that often takes place outside the realm of the production of the instrumental portion of a song. There is a separation in the production space of these two practices as they feature different problems to address and require different skills. This tends to create specialists that focus on one discipline or the other. After lyrics are prepared, vocal production can require renting a suitable recording space, sourcing the correct gear, solving logistical issues surrounding singer availability. As songs evolve, more sessions are booked to address changes that must take place in the vocal recording, so the process often is repeated. A change in production can require a change in the vocal work, and reversed. The separate nature of these two practices greatly increases composition and production time, as one process awaits the completion of the other.

I'm not a singer, my specialization has been music composition and production. I have not been able to develop the abilities to hit desired notes when singing. Most importantly, I do not find my voice to have an interesting texture. It is difficult to write lyrical music if you can't sing as it's hard to get a sense of how lyrics sound in context. Even if training resulted in a decent voice, I think about my relationship to the guitar and its historical place in my song writing. If I could take a step back, analyze the chords, melody, and program vocals, would I do better work, similar to what happened with

guitar? Why get held back by the limitations of my own physical ability if it was possible to get a realistic and decent topline in the digital space.

As a music maker using digital tools to compose and produce music, I was constantly on the lookout for new developments that could support my work. I used the internet and wide contact networks in the music industry to identify new products and approaches and then tried out ones that seemed promising – basically an applied research approach. This was how I approached my search for a singing voice tool or system that I could incorporate into my music making. Did the tool meet my needs? Did it have ease of use? Did the output sound like a natural voice? And most of all, would it help me make the type of music I wanted to make?

Context

The following context chapter provides an overview of speech synthesis and singing voice synthesis (SVS) development, a review of technologies developed to date and the applicability to my music making need for a singing voice synthesizer tool. It looks at current speech and singing synthesis technology and the drawbacks and limitations of the technologies with regard to music making applications.

Speech Synthesis Overview

Speech synthesis is the artificial production of human speech. Today, the majority of speech synthesis happens within a computer system that takes text as an input, and outputs synthesized speech. This is known as a Text-to-Speech system (TTS).

There have been two common goals throughout the history of speech synthesis development; intelligibility and naturalness: intelligibility can be judged simply by whether words could be correctly recognised; naturalness is less defined. For the purpose of this research project, the concept of ‘naturalness’ was considered in terms of whether it sounds like a live human, or is noticeably robotic in texture, and where it sits on this scale. Essentially, does this generated speech sound more like a robot, or more like a human?

Many techniques have been developed to create a natural and convincing synthetic speech engine. Often, these techniques have been based on the concatenative method, a process where a large amount of speech is recorded from a single speaker to build a database of smaller speech fragments that can be reconstructed. Pre-recorded fragments of words are strung together, and then these smaller units are recombined to form speech. They are usually limited to one speaker, and require a great deal of memory to store the necessary data for speech synthesis. Constructing a speaker out of these smaller audio units makes it difficult to modify the voice without recording a whole new database. In order to create a second speaker or speaker style, the process must

be repeated (Oloko-Oba et al. 2016). A different technique, called parametric synthesis, relies on the analysis and modeling of the spectral characteristics of speech recordings. The parameters learned in the analysis are used to reconstruct speech by simulating the vocal tract of human beings using a parametric physical model (Zen, Tokuda, and Black, 2009). Concatenative speech generation typically produces more natural sounding speech, but its effectiveness is limited by the size and quality of the recorded database. Parametric synthesis allows for greater flexibility in speaking style by modifying parameters of the model, but generally is less natural.

Speech synthesis systems often see a trade-off between realism and malleability. WaveNet, presented in 2016, was a major breakthrough in the area of speech synthesis and involves using neural networks to model the human voice. WaveNet demonstrated an ability to produce natural-sounding synthesized speech that outperformed previous methods, by directly modelling waveforms. Directly modelling waveforms is difficult, for instance a standard clip of audio can contain tens of thousands of samples per second. WaveNet presented a solution by using techniques found in Pixel-CNN, which is a model that can be conditioned on an image to generate similar images. It generates pixel-by-pixel by looking at data from previously generated pixels. Pixel-CNN uses dilated convolutional neural networks, which enables making increasingly large skips in the data to generate the larger picture. This method of skipping data makes direct modeling on waveforms possible. While WaveNet outperforms older methods in terms of naturalness, it is computationally expensive and extremely slow to train, so not very malleable. WaveNet as a standalone framework for TTS purposes may be outdated already, but it's a useful starting point to build a foundation of TTS and audio synthesis knowledge (Oord et al. 2016).

A solution to the resource intensive process of WaveNet based audio synthesis was proposed in 2017, the original Tacotron. This model takes characters as input and outputs raw spectrograms, which are synthesized using the Griffin-Lim algorithm. The Griffin-Lim algorithm greatly reduces the time to generate over WaveNet, however it produces undesirable artifacts and lower quality audio (Wang et al, 2017). Tacotron

attempted to solve the speed issues of the original proposed WaveNet, but at the cost of quality. Tacotron 2, presented in 2017, improved on the quality of the original Tacotron and significantly reduced the various costs of the original WaveNet architecture. Tacotron 2 took the best parts of the original Tacotron, mapping a sequence of letters to a sequence of features that encode the audio, which are then synthesized by a modified WaveNet vocoder. The results of Tacotron 2's speech generation were of impressive quality and featured a lower price tag in terms of computational cost and time to train. These qualities made this model especially attractive to experiment with. Online forums filled with hobbyists and professional developers alike have grown over the years around speech synthesis. Out of the available open source implementations of neural speech synthesis systems, I have found Tacotron 2 to be the most widely understood and discussed, greatly reducing the challenge of working with a complex machine learning framework.

Singing Synthesis Overview

In this section, I consider three different types of singing synthesizers; Vocaloid, DeepSinger and a neural parametric singing synthesis system based on the WaveNet architecture.

Vocaloid is a commercial singing synthesizer in which the underlying technology has gone through many iterations (Bonada et al. 2017). It has improved greatly over the years, but its underlying technology is not that much different from diphone speech synthesizers: Vocaloid is based on sample concatenation, where the singer library is a database of samples (diaphones) extracted from real human singing. Vocaloid's system consists of a synthesis engine, a score editor and the singer databases. The synthesis engine is used in the selection, modification and concatenation of the sequence of diphones. The singer database contains the diphone samples and related analysis which are recorded at three pitch ranges (low, medium, high). The problem in concatenating samples is that the samples are recorded in different pitches and different phonetic contexts. Therefore, after a pitch is selected, time stretching and fundamental

frequency transposition have to be applied to match the selected note. So, building a vocaloid database requires recording a large store of audio, making it difficult to build new singer profiles, as well as pitch manipulation which leads to mechanical sounding output depending on the distance the pitch is shifted from the original recording (Kenmochi and Ohshita, 2007). Vocaloid's need for the extensive fine tuning of available parameters to achieve natural performances makes it a difficult and time-consuming system if your goal is realism. Besides the Vocaloid often resulting in unnatural sounding singing, it is a closed system. Users are not able to make custom voice banks and, therefore, are limited to those provided by the company. This means that if Vocaloid is being used in a production, the stylistic qualities of the synthesized vocals are defined by what a single company chooses to produce and make available. Keeping up with new and emerging vocal styles is not a feature of the Vocaloid environment. The majority of Vocaloid's success has been seen through the marketing of digital avatars, drawn and animated in the stylistic realm of Japanese anime. Their success is highly related to synthetic characters rather than professional music production, so keeping up with trends in pop and underground music styles most likely is not a key objective of their business. To summarize, it is difficult to achieve naturalness with Vocaloid and even more importantly, users are limited by what voice banks are offered for sale by Vocaloid.

DeepSinger is a state-of-the-art system that proves singing data could be mined and implemented in an SVS system without much human interaction. DeepSinger focuses on generating singing voice synthesis after vocal separation from a complete track. Instead of bringing a singer in to sing for hours to build a dataset, the DeepSinger system pulls complete songs from the internet, separates the vocals and the underlying instrumental then automatically transcribes the lyrics for the isolated vocal track. DeepSinger uses an open source ML vocal separation tool Spleeter, which I use in my own music production pipeline. When a topline pitch is sent to me, it is often in the form of a single audio file including basic production. I use the Spleeter script to create stems, enabling experimentation with a song before agreeing to a production contract. In my experience, while Spleeter is a useful tool for rough drafts, artifacts and the

underlying original instrumental are always present in the extracted vocals, resulting in an unusual and unnatural sound. The output of DeepSinger retains much of the unnatural sound that Spleeter creates through multi-track separation of data based on what high quality recordings can be sourced. Therefore, this is currently not a system to build SVS systems on specific high-quality voices. This is not consistent with my goal of creating a system that can be used to easily create digital versions of specific vocal styles, preferably built with limited but high-quality recordings. Furthermore, DeepSinger's method of mining available data from the internet is an advanced engineering solution to the data problem, but is this method actually feasible in the field of music production? The methods used raise questions of legality related to the source material. Making digital clones of voices without permission is likely crossing lines in the area of copyright infringement. See Appendix on legal issues.

In 2017, Blaauw and Bonada presented a neural parametric singing synthesis system based on the WaveNet architecture. Their research group, Voctro Labs, has also been a key collaborator with Vocaloid, and has had many breakthroughs in the field of audio and voice technologies. The objective of their work was to address issues of flexibility and scalability of concatenative synthesis, the method used by popular singing synthesizers like Vocaloid. The system developed is based on separate but interconnected models that learn pitch, phonetic timing and timbre from a dataset of songs. They found this to be a solution for the difficulties surrounding creating a large singing dataset as it required less training data. The system uses two types of singing data - natural singing and, what they call, pseudo singing. Their English pseudo singing dataset consists of 35 minutes of short sentences which were sung at a single pitch and an approximately constant cadence. While their system simplifies the problem of synthesizing any melody, there appears to be degradation in sound quality during changes in pitch. The results from their English examples are expressive, but fairly robotic and tinny. It feels like it is missing harmonic information and the results are highly unnatural with a prominent nasal-like quality. The Spanish and Japanese demos seem to be of moderately higher quality, but it is difficult to make a qualitative analysis on an SVS system in a language I do not understand. The audio demos provided are

demonstrations of impressive engineering, which is their intention, but the results are below standards of naturalness for my own use. While independently modelling pitch and timbre allows adjusting pitch of the generated vocals to match a target melody, the resulting sound seems to be a good representation of the tradeoff between malleability and naturalness. Moreover, if a resulting vocal tool were a closed system where voices were solely provided by the developers, possibly more numerous due to the underlying technology facilitating the creation of new voices, it would not meet my interest in a flexible system that allows user created voice datasets.

Context Discussion

The breakthroughs in neural speech synthesis as seen in WaveNet, leading to Tacotron 2, prove that highly realistic synthetic speech can be produced. However, singing tools, using established methods such as Vocaloid or even cutting-edge additions into the field, still leave much to be desired. From a personal analysis of the outputs, they contain qualities that are unmistakably non-human, and the complexity of the technologies makes it difficult to build new voices. While it is possible to output realistic singing from Vocaloid with the investment of a lot of effort and time, Vocaloid does not give the option to create custom voices and I assume it would be quite difficult to create an effective one, judging by the underlying technology. Virtual instruments can be realistic substitutions for the real thing. Almost every instrument is available and can be programmed into any style. Building an instrument oneself is fairly simple as well: Record multiple notes and desired articulation styles, then use them within a multi-sample player within a DAW (Digital Audio Workstation). Building new virtual instruments is not a complicated task. They can be shared between music makers, and there is a large resource of sounds that are kept up to date in terms of constantly evolving musical tastes. However, there are not many options for synthetic vocals, and at the highest level of vocal synthesis, the results are subpar. Vocal synthesis has not reached this level of rapid evolution. Presets and styles are based on what engineers develop without artist or cultural input.

It’s important to note that a lot of these additions to the field of SVS (besides Vocaloid) are mainly research focused. There is ground-breaking work happening within these inventions and innovations but they do not often work in the context of actual music production. Many are presented as possibilities, to be further developed and used in a future product. It would seem that the dream of having a bank of singing voice styles is a long way off.

However, looking at the results from open source implementations of the top speech synthesis tools, the tools needed might already be in place. Can creative design strategies be employed within the existing advancements of speech synthesis to realize a framework for easily adaptable, customizable, and natural synthetic singing? I aim to create a synthetic vocalist system that can be used in my workflow, where non-engineers can create voice banks to keep up to date with trends, therefore can output music that would be of interest to my peers and networks. While these SVS projects are great feats of engineering, the researchers most likely do not share my specific goal, which is to create a tool that can help me make music that would succeed within the current landscape of contemporary pop.

Methodologies

The methodologies used to develop this thesis project included user-centred design with myself as the primary user, research through intuitive design and iterative prototyping to realize the thesis product, a singing voice synthesis system based on open-source speech synthesis technology.

Research-through-Design

My design research approach has been essentially what Christopher Frayling describes as “research for art and design”(5). As a music maker working in a new and rapidly evolving medium and practice, I have had to continually research the latest technologies, teach myself the digital skills needed to exploit the tools available, modify tools to fit my needs, and develop new tools as needed. For me, Frayling’s description of “Research for art and design…where the end product is an artefact…in the sense of visual or iconic or imagistic communication” (5) seems closest to describing my approach. It includes the elements identified by Frayling (1993) of materials research, i.e., into the related technologies of speech and singing synthesis, and action research using an iterative prototyping process to design and realize the project. It takes an intuitive design approach in that the design process is guided by my intuitive judgement based on knowledge acquired through extensive experience in music making and the use of digital music making technologies. Dijkstra et al (2012) describe ‘intuition as a process of thinking, the input of which is mostly knowledge primarily acquired via associative learning’.

Iterative Prototyping

Key to realizing the project was the use of iterative prototyping as my key design methodology throughout the development process and testing each prototype myself, as the key user, asking the question, to what extent does the result meet my needs as a music maker?

Designing within a new technology presents problems in itself. The limits of the technology is not always fully understood, or cannot be fully realized by the resources at hand. Therefore the process of developing solutions within an evolving and complex field required an iterative prototyping approach. Even in engineering problems, the design of systems is rarely accomplished exclusively by applying fundamental scientific principles and requires some use of experimentation (Dym, 106). Having an idea about how to address my need as a key user and extensive experience in the digital music sphere, the most direct approach to designing a solution was to design a series of prototype models with each model building on the learnings from the previous model. Prototyping is used as “a tool for thinking...a process of thinking taking place through the act of making” (Mulder, 2018). It is effectively applying Frayling’s “action research” process, a “series of practical experiments” effectively researching through design (5).

Project Development

The next section details the research and development of a singing voice synthesis system. It introduces the research work with a description of the research objective, the known data issues and the research approach. I then describe my research process, followed by a description of the research results and a discussion of those results.

Research Introduction

Research Objective:

> To create an effective and convincing singing voice synthesis system with a simple process for building new voice styles that could be used by music makers to create new toplines independent of a live singer.

In the online communities based around speech synthesis, hobbyists are able to create realistic speech using open source systems like Tacotron 2. Natural speech is proven to be a possibility with these open-source implementations. While it can take weeks to train a Tacotron 2 model, effective training can generate realistic speech in seconds. Machine Learning based speech synthesis faces the usual challenges surrounding machine learning problems: The accuracy of a model's prediction is directly related to the quality and size of the dataset. To effectively train a speech synthesis model, a large dataset of recorded speech must be used. For example, the open source LJSpeech dataset, a popular resource, contains 13,100 short audio clips, including transcriptions, of a single speaker that varies in length from 1 to 10 seconds. The total length is approximately 24 hours (Ito & Johnson, 2017). It is clear that the creation of speech datasets is a laborious task that requires many hours of recording, large storage capacity and computing power to manipulate.

The more data available of a single speaker, the higher quality the generated output will be in terms of fidelity and ability to replicate the original speaker. While there are many sources for single speaker recordings like the LJspeech dataset in the public domain, robust single singer datasets are non-existent. The isophonics Singing Voice Dataset provided by the Centre for Digital Music looks to solve one of the main problems in singing research - the lack of data in the form of unaccompanied singing. However, there are two issues with this dataset. First, at a total length of recorded audio of just over two hours, it is a small dataset in terms of training a speech synthesis model. Second, and even more limiting considering my own purposes, the greater part of the dataset is Chinese opera music which has limited application to creating English contemporary music. Another dataset is the (C4DM) VocalSet: A Singing Voice Dataset. This is a singing voice dataset of a capella singing that captures a wide range of styles and consists only of recorded vowels. This dataset could be used to train machine learning models to separate mixtures of multiple singing voices, or understand the difference between different singing modulation techniques (vibrato, straight, trill, etc.) (Wilkins et al., 2018). This dataset just contains sung vowels, therefore words would need to be constructed from these vowels. Even if a neural synthesis model could replicate these vowels and techniques flawlessly, they would have to be strung together in a concatenative fashion in the process. It is evident that unaccompanied singing data is a limited resource, and the solution is to create an entirely original corpus from scratch.

I developed three iterations of a prototype design for a singing synthesizer system. At each iteration, the experiments were small in scope. This was to ensure time was not wasted when considering the resource heavy nature of training neural networks to reproduce audio.

The first step was evaluating the potential of the technological innovation, through the appropriation of original engineering to fit my desired outcome. At each iterative stage of prototype development, an analysis took place of limitations and/or possibilities presented. This is seen in the initial prototype stage by experimenting with regular

speech models to understand what would be the most essential vocal music parameters needed within the system.

The most important factor that enabled successful work within the solution space was a full understanding of the problem space from the user's perspective, and especially the art form's cultural context. This was fully realized during the second prototyping stage and is seen in the analysis of what a user might actually want out of a vocal synthesizer. The methodologies in vocal synthesis papers mostly detail strictly engineering related problems and methods for approaching their solutions. By analyzing the actual use case, the system architecture was simplified and development was less complex. Understanding the desired outcome from a user and cultural perspective was instrumental in terms of facilitating the appropriation of their original engineering to fit my desired outcome.

Prototype 1

First Prototype Question: What is needed as output from a standard speech synthesis model in order to create vocal music?

The scale of audio data needed to train a model for effective speech synthesis means that errors in the creation of a vocal dataset can be expensive in terms of resources spent (hiring a vocalist, time spent recording and prepping data, cost to train models). To mitigate the risk of spending resources unnecessarily, it is important to decide what is actually needed as output from a model in order to create convincing vocal music. To explore this question, I began by attempting to create a topline with regular speech from a previously trained model.

First Prototype: Create music using a non-singing synthetic voice.

The first step in this process was finding an average pitch of the synthetic speech then tuning complete phrases to this pitch. With this step, I created a single note baseline that allows me to pitch shift the audio around to easily create a melody. For example, if the phrase was tuned to A, a word could be pitched up three semitones to hit C. This handled manipulating the audio melodically, but singing also has a rhythmic component. Adjusting the playback speed with Ableton's time stretching algorithm was used to affect the rhythmic portion of the topline. The results sounded satisfactory through the application of a deep dive into vocal production techniques using Melodyne (a tuning application), vocoding, pitch and time shifting algorithms built into Ableton software. The more these techniques were applied, the less natural the voice sounded. However, while the voice was fairly robotic from extensive digital pitch manipulation, I found the music acceptable to my personal tastes.

Prototype 1, "Hold On", Nov 2019
https://vimeo.com/526958419

In conclusion, simple pitch and rhythm adjustments created a representation of singing, and if a model could generate output audio that didn't require extensive pitch or time stretching, then the generated audio should be able to retain naturalness.

Prototype 2

Second Prototype Question: What parameters of singing will be necessary to create modern popular music? How far back can singing be stripped to a convincing topline?

The initial output from a speech synthesizer does not have musical intention, it must be creatively shaped to fit rhythmic and melodic qualities of singing. The question to address is whether a SVS system needs multiple highly adjustable parameters that simulate techniques used in complex vocal performances. The results of the first prototype show that if a model could generate audio that didn't require pitch or time stretching, the generated audio should be able to retain naturalness. While still pitch

shifting from a single base frequency, the synthesis would have a cleaner result if the output did not need the initial tuning to a single note. However, since there were no datasets available of singing at a single pitch, I would have to create a new dataset.

I anticipated three challenges to building a neural synthetic singing voice system (SVS):

1. creating a dataset, featuring notes and timing,
2. understanding the complexity of artificial speech synthesis, and
3. the resource intensive nature of time and computational cost of training models involved in neural audio synthesis

The datasets involved in speech synthesis contain hours of speech, up to 24 as seen in the commonly used LJspeech dataset. Achieving optimal dataset size is difficult, but there are the added complications of pitch and duration of lyrics being sung. Earlier experiments show that basic outputs required to create vocal music are pitch and duration. Pitch stretching regular speech could create interesting ‘sung’ pieces. Part of that process was finding the average frequency of the output speech and tuning the complete phrase to that pitch. To skip this step, a prototype dataset could be made by singing at a single pitch. Systems like Tacotron 2 have an easier time modeling datasets that are uniform in tone and pitch than those with high divergences such as multi-speaker datasets or single speakers doing characters (Olney 2019). If training on a single note at a consistent tempo, I assumed that the models would learn quickly and require less data. If the models need less data, then perhaps it would be easy to build a range of registers and pitches. Single pitch datasets create an opportunity to employ a technique in speech synthesis known as voice cloning. Training TTS models from scratch is an intensive process, and often there is not enough data as seen in the 24 hours of the LJspeech dataset. Voice cloning uses data learned from a large speech corpus, then the learned parameters are adapted to the smaller dataset. If the model does not need to consider pitch and rhythm variation, then continuing training from a model created using a large speech corpus is possible. Combining voice cloning with single pitch singing provides a solution to the hours of training data usually needed in TTS systems. Tacotron 2 is state of the art for TTS, so success is more likely if building

the dataset to fit Tacotron 2 rather than attempting to manipulate the technology to fit the data.

A dataset containing a voice singing at a single pitch and a uniform rhythm could potentially solve all challenges by:
1. reducing the required data
2. adapting the data to fit Tacotron 2, rather than the other way around.

Second Prototype: To develop a prototype using a dataset of singing at a single pitch.

In the first prototype, it was seen that pitch and rhythm adjustments could communicate lyrics in a musical style. The question to address at this stage is whether a SVS system needs multiple highly adjustable parameters that simulate techniques used in complex vocal performances. Personal experience has shown that there is currently a greater emphasis on rhythmic vocals with simple melodic forms, over virtuosic performances. Looking at what music analytics indicate are the dominant features of contemporary music can put this theory in context. MRC Data is the most comprehensive global provider of data and analytics to the entertainment and music industry and consumers. According to their metrics, as of 2017, R&B/Hip-Hop became the most dominant music genres, with seven of the top 10 most-consumed albums categorized as such. (Nielson 2017). The year 2020 saw album sales vary in genre, but digital song consumption was dominated by R&B and Rap, which also led in terms of total volume of consumption at 28.2% vs. Rock at 19.5% (MRC Data). Modern Rap sees the use of rhythmic melodic hooks more than its lyrical focused early days. This success might be affecting the vocal stylings of other genres within pop music, or simply represents the average listeners' topline style preferences. One of the most popular songs of 2019, Billie Eilish's 'Bad Guy', is a song built around timbre and rhythm, a performance that can be described as a soft rhythmic whisper. Records like 'Bad Guy' are examples of the importance of a catchy swagger over demonstrations of vocal ability in modern popular music.[1] Unlike

[1] https://www.nielsen.com/wp-content/uploads/sites/3/2019/04/2017-year-end-music-report-us.pdf

more traditional forms of singing such as opera or musicals with far more complicated parameters, the number of style inflections has narrowed with the popularity of contemporary rhythmic music. So, based on trends in popular music, multiple, highly adjustable parameters may not be necessary to create modern popular music. There may only be a need for controlling pitch and duration.

The conclusions drawn from the current state of popular music is that we do not need an extreme range of vocal articulations and pitches. An interesting timbre and a few notes might just be enough. While there is not much in the way of single voice singing datasets as a resource, recordings of singing at a single pitch is non-existent. Either way it was assumed that a new dataset would need to be created, and the theory is that single pitches will be easier to train. A constant rhythm could be set by having the onset of a consonant or vowel commence at the beginning of each beat of a metronome.

To test this theory out, I recruited an artist who is a frequent musical collaborator, and also a voice actor. The plan was to record as much audio at a single pitch, then train the Tacotron 2 (T2) on the smallest amount of data possible. If the model was not able to learn at a small data size, more would be added. This test would set a guide for future pitches and registers. The recording was done out-of-studio using a decent microphone and acoustic treatment blankets. While we both have access to professional studios, the amount of experimentation needed made home recording the economical choice. The dataset created from these sessions was organized into audio files with the length varying between two to ten seconds per spoken line which is found to be optimal for T2. Since T2 is set up to train on regular speech with great results, I molded the concept of singing to fit into the T2 setup. The process is the same as training on regular speech, but the output when combined with other trained T2 models, could be put together to form singing. So, the approach was to first train a model on speech-like singing at a single pitch, then put together multiple models to form a topline. Throughout training,

---

https://www.musicbusinessworldwide.com/files/2021/01/MRC_Billboard_YEAR_END_2020_US-Final.pdf

parameters of T2 were fine tuned, a mostly trial and error process - train for days, check output, experiment with parameters. After multiple iterations of the training process, lyrics could be given as input to the trained model, which would generate the line sung at a single pitch. While producing the melodic topline required pitch stretching, it could create toplines that I found satisfactory. This process requires close to 2% of the LJspeech dataset, which also means a more complete vocal range could be built without extensive training data. By simplifying the needed output to basic parameters, the majority of the time invested is spent on training the models, and solves the issue of building a large dataset.

Prototype 2, “Maya Training Samples”, March, 2020
https://vimeo.com/526958419

The topline melody and rhythm took about 2 hours to work out over a basic instrumental. Creating adjustments to the lyrics, melody, and standard music was straightforward as it was a single contained workflow without the need to book more recording sessions, or being confined to how a topline is originally presented. The results proved that utilizing novel design strategies within the current state of speech synthesis could create a unique solution to synthetic singing. For the first time in my practice, malleable instrumental and vocal production was taking place in the same space. This iteration saw a single pitch model trained with limited data. The output still required extensive pitch manipulation to replicate the melodic forms of singing, which introduces artificial qualities to a voice. If a range of pitch models were available, this step would not be needed.

Prototype 3

Third Prototype Question: Once the individual pitch models are integrated into a system, does the system meet my needs for a synthetic vocal solution?

The first two prototypes helped develop and test a novel approach to create synthetic, singing vocals within synthetic speech technology. The strategy is to use a small dataset of single pitch singing combined with parameters learned from regular speech. The second prototype also helped create guidelines for the minimum data required from the recording stage. The next step was to record multiple pitches, then train separate models on each pitch. This would provide a range of notes without the need to adjust pitch within an audio editor. Creating interesting rhythms would require adjusting the playback rate of moments of singing as the original data features a syllabic rhythm that is replicated in the output. Manipulating the playback rate of samples can degrade quality, although it is only noticeable if used to a high degree.

Third Prototype: Assemble various models in a complete SVS system.

The system is set up from training separate models on individual pitches, therefore the first step is inputting lyrics into each pitch model. The output will be the line sung at each selected pitch, all following the rhythm of the original recording. This output is then put into Ableton, where every pitch is placed on a separate track. Once the audio is formatted, individual syllables can be activated or muted to create a melody. The pause structure affects the sound of the speech in the sense that the beginning of a sentence generally contains different inflections from the end.

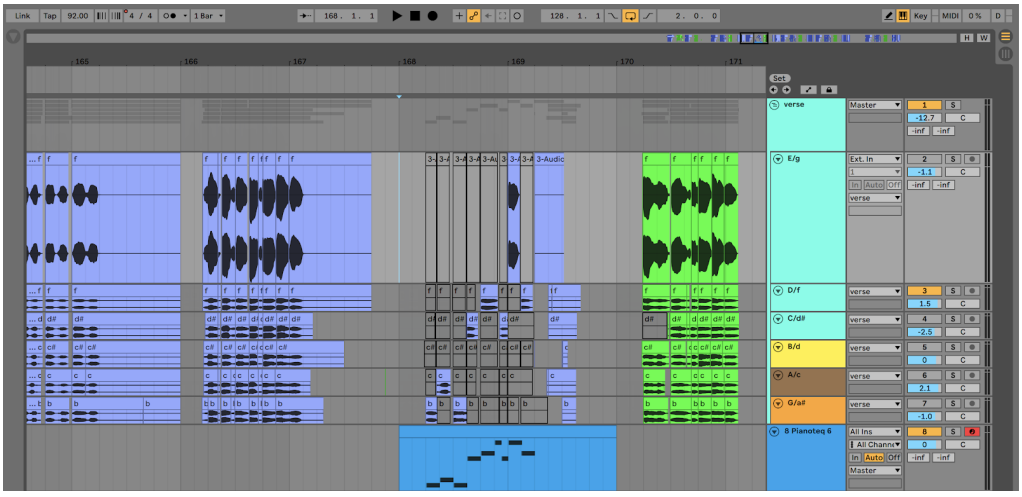

Figure 1 - The diagram shows working with multiple instances of single pitch singing within Ableton Live. Horizontal layers represent separate pitches, while lyrics and timing are identical. At the bottom layer is a midi melody planning out the melodic form of the topline. Vocal syllables are represented by blocks above the midi layer. Blocks that are opaque are audible, while transparent ones are muted.

In the above example I have worked out a melody in MIDI notation, then edited the individual tracks to match. From lyric input to the point where the melody is being experimented with took no longer than 30 minutes. The immediate result is far from natural, and this is common for almost all live instruments replicated in digital space that have not undergone processing. As discussed earlier, strategies to make guitar sound natural involve incorporating human elements like imperfect timing. Understanding variables of live performance helps effectively adjust digital instruments to sound convincingly real. Even before the generated singing is manipulated from its current rigid state, it already proves to be an effective tool for building a topline concept. Melodies, lyrics and underlying composition can now be developed simultaneously and swiftly, providing efficient access to a process that a complete lack of vocal talent has kept at a distance.

The following is a sample of a song made with the third prototype assemblage. Modern vocal production techniques were applied to the generated vocals, such as equalization, compression and reverb. What is immediately evident is the generated voice unmistakably replicates the vocal qualities of the original singer and dataset prosody style. The voice sounds like the original singer, but the rigid prosody reproduced from the data has to be manually adjusted to resemble a live singer. With some experience I found the time required for the adjustments similar to the adjustments needed to make a digital instrument sound convincing. With pitch, timing and lyrics end-to-end accessible, I landed on a topline that fit my tastes. Whether or not the song would be enjoyed from an outside perspective is not a measure of success for the project. I was able to

single-handedly create music that met a self defined aesthetic target, and this is what's important.

Prototype 3, "Bad Habits Demo Sample", February, 2021
https://vimeo.com/526958419

While the system has proven to be an effective tool to build topline ideas, I aim to see how close this prototype iteration can come to a real vocal performance. The main areas that decrease the output's naturalness are found in the strategies used to build the original dataset. Combining voice cloning and single pitch singing solves data requirements but creates variances between pitches that one would not usually hear during a live vocal performance. During live singing, the delivery of each lyric is affected by the previous pitch, pronunciation and annunciation. Recording at only a single pitch will flatten these variances, which can lead to an unnatural prosody. At my request, the collaborating vocalist made an effort to not be overly expressive in pitch and clearly enunciate vowels and consonants. This was done out of caution so input data would be as uniform and clear as possible to help the training process. This causes the original vocal delivery to feature spaces and heavy annunciation between syllables that is much more pronounced than a live singing. This rigid pitch form and strict delivery is the aspect of the system I find most unnatural. While these issues are not ideal, the silver lining is that it is accurately replicating elements built into the original data, therefore can be addressed in a future dataset iteration. A comprehensive and research backed design strategy is in place which shows that it's possible to build the required data in a single day. In terms of unnatural prosody, focusing on connecting syllables within whole words and a less pronounced delivery will help solve this. A future dataset that incorporated more natural vocal stylings would improve the output, but expressive singing would likely be more difficult for a model to understand therefore increasing data requirements.

Production Notes:

Unlike other synthetic singing solutions, the act of changing pitch of a sung vocal is non destructive in the sense that it does not require digital pitch manipulation, resulting in unnatural artifacts and degrading audio quality. This current version took time between recording sessions to make sure each pitch could be learned by Tacotron 2 with the proven strategy. Future singers will be able to complete a full dataset within a single session. The time to complete a vocal session with a live vocalist, finalize lyrics, record and re-record, sign a royalty split agreement can be years. A simple song with this framework can be a matter of hours, or less, depending on the amount of tinkering is done. The issue of setting up a re-recording session if a part of a vocal take needs to be fixed is also easily solved, as any note available through the models can be generated. If a word isn't working, that word can be swapped for a new output. A topline can be created all within the computer.

There are minor issues in terms of pronunciation and intelligibility, but most are manageable as they add only a minor amount of tweaking to the process. Generating the correct pronunciation sometimes took experimentation, and each pitch model had its own pronunciation issues. If the data did not contain similar speech, it had difficulty generating it, therefore, I had to edit small units such as vowels together to form a complete word. Mispronunciation can be fixed by substituting homophones, such as substituting "our" in for "hour". Occasionally words could be broken up by their syllables. If a model had trouble pronouncing "baby", breaking up the word can help such as inputting "bay" and "bee". Certain pitches proved to be more difficult. It is difficult to say why a certain model's quality is lower in comparison to the other pitches trained. It could be because the sound of the voice at that pitch is not like the underlying original model resulting in recording and processing errors. Some of these issues were addressed by retraining the models and adjusting parameters. If the data did not contain similar speech, it had difficulty generating it, therefore, I had to edit small units of generated speech together to form a complete word.

Project Results

Using a combination of advances in neural speech synthesis, a specially created singing voice training database and a background of experience in the digital music industry, I have developed a solution to vocal singing synthesis that meets my needs as a contemporary music maker. The system is built around a training voice data set that lies somewhere between singing and regular speech which can be fit to the Tacotron 2 speech synthesizer to produce voice output. The voice output can then be input to Ableton where every pitch is placed on a separate track. Some formatting is required in terms of timing, as while the original source audio was recorded to a single BPM, the singer was not always accurate. Once the audio is formatted, individual syllables can be activated or muted to create a melody. The system is made up of recording 7 different pitches for around 30 minutes each.The system uses such a small amount of data because it employs voice cloning from models that were extensively trained on regular speech. It continues training from learned parameters from large regular speech datasets, and applies parameters to the smaller new dataset. Because the new dataset is highly uniform in pitch and rhythm, the model is able to quickly understand the new data. Single pitch, uniform cadence singing can be described as a human imitating a robot. Interestingly enough, it turned out that the data needed to be less lifelike for the final product to achieve realism.

The output will still require rhythmic adjustments to create an interesting topline, and stretching or shrinking a syllable can degrade the audio fidelity depending on the amount. While working on this project, other developments in speech synthesis occurred that dealt with controlling the speed at which a line of generated speech is spoken. A possible solution is the open source ForwardTacotron, which is a modification of Tacotron 2 that allows control of the speed of the generated speech. Developing toplines in the same digital space as instrumental production has joined two disciplines with traditionally distinct workflows within my practice. The methods used set up a simple framework for the creation of new voice models, which opens the possibility for a wide range of vocal styles.

Discussion

The system that I've assembled has an almost primitive feel when compared with the work of the top researchers in the field of singing synthesis. It uses a brutally simple strategy within a complex field to achieve a specific goal. I believe that building a singing voice system, based on such a simple strategy, is key to successfully bringing singing into the digital realm. Whether due to the difficulty in creation, or a company's desire for product control, a closed system where the presets are solely provided by the sofware's developers is too restrictive for artistic creation. This limitation has caused the current leading synthesizer Vocaloid to fail outside of its niche anime related market. It has no ability to adapt to current vocal trends, such as register styles like vocal fry or the evolving articulatory styles seen in modern Rap. Not all genres are defined by vocal style, but as the singer is often the focal point of contemporary popular music, it is an extremely potent force. Modern popular Country is a good example of how vocal style is often the key driver of the conceptualization of a song's persona. The production of the 2017 Country hit 'Meant to Be' by Bhebe Rhexa and Florida Georgia Line sees generic piano and electronic drums popularized by Rap subgenres. Bhebe Rhexa's vocals on 'Meant to Be' could be supplanted over any pop song, while Florida Georgia Line provides a distinct Country twang vocal style. Modern Country records often see underlying production normally associated with R&B, Rap, and standard Pop, yet their toplines push them well into the country realm. Vocal style often defines modern rappers like Migos or Future, who use an articulatory style involving creative and sometimes incoherent annunciation. Rather than pure lyricism, there is emphasis on the production, vocal effects, melody and overall swagger within the delivery. Unlike the country twang, this is a recent invention. The vocal style has given way to a new genre classification, known as 'Soundcloud Rap' or 'Mumble Rap', alluding to its, at times, unclear delivery. A system built on standard articulation would not be able to replicate this trending vocal style, yet it is a defining aspect of some of the most popular artists that dominate charts. Vocal systems are often built on the concept of malleability in a parametric sense, but

not culturally. Creative applications, to be effective in ever-evolving respective creative mediums, need to be malleable by the users themselves. Vocal style is the most important active parameter for experimentation in modern music. If future vocal synthesis systems are to succeed they must be agile enough to adapt to trends and the sound must not be solely defined by the engineers behind the technology. The sound has to be driven by the artistic community.

Conclusion

The objective of the research project was to build a convincing, natural sounding singing voice synthesizer that music makers could use to create new toplines and the system featured a simple solution to build new voice banks. My overall research question for this project was how might I create an effective synthetic vocal tool using current speech synthesis technology that is beneficial to my practice? My research strategy was to design a solution through an iterative prototyping methodology informed by my extensive and in-depth user background in technical audio knowledge and practice-based music production. In retrospect, I would describe my approach as a research for design strategy, what Frayling (5) describes as one “where the end product is...embodied in the artefact, where the goal is not primarily communicable knowledge in the sense of verbal communication, but in the sense of visual or iconic or imagistic communication.” In my case, the “artefact” is a new and practical approach to a singing synthesis system. Through prototyping, theories were developed that were used to build an effective digital vocal solution.

I set out to solve a long running problem throughout my career based solely on my own design intuition and artistic sensibilities. Though I did not have a technical background in machine learning, I have developed skills with computer based audio tools including creating and adapting specific tools to meet my needs. The success of my project stems from applying a research through design methodology guided by intuition grounded in an extensive background in technical and artistic audio design. A deeper understanding of the voice and its place in contemporary music from a user perspective led to the development of novel solutions.

The production process addressed technical problems and theories for the development of creative software. The approach also led to another lesson learned, that the designer should not be intimidated by the complexity of the problem space in a familiar field. A complex problem often has multiple solutions, and if the medium is familiar they may have a fresh approach to the problem.

Further research

Any unnatural results from the assembled system were a result of the unusual prosody of the original data. The current framework could be improved by building new data sets that experiment with a less strict delivery. While there are paths to improving future iterations of this project, the current results also have implications for the field of singing synthesis from a user perspective. Reproducing the quantitative properties of vocal performance is not necessarily what will make a tool valuable to producers. In my practice, and in much of modern music, the aesthetic experience of a voice is more important than technical skill.

Bibliography

"An Act to enact the Consumer Privacy Protection Act and the Personal Information and Data Protection Tribunal Act and to make consequential and related amendments to other Acts". *Bill C-11 First Reading*. Minister of Innovation, Science and Industry. House of Commons, Canada. November 17, 2020.
https://parl.ca/DocumentViewer/en/43-2/bill/C-11/first-reading

Balyan, Archana et al. "Speech Synthesis: A Review." *International journal of engineering research and technology* 2 (2013): n. pag.

Biing Hwang Juang et al. "Digital Speech Processing", Editor(s): Robert A. Meyers, *Encyclopedia of Physical Science and Technology (Third Edition),* Academic Press, 2003, Pages 485-500,
ISBN 9780122274107.
https://doi.org/10.1016/B0-12-227410-5/00178-2.
(https://www.sciencedirect.com/science/article/pii/B0122274105001782)

Blaauw, Merlijn, and Jordi Bonada. "A Neural Parametric Singing Synthesizer". *Music Technology Group*. Cornell University. 2017. arXiv:1704.03809 **[cs.SD]**
https://arxiv.org/pdf/1704.03809.pdf

Dijkstra et al. "Deliberation Versus Intuition: Decomposing the Role of Expertise in Judgment and Decision Making". Journal of Behavioral Decision Making, J. Behav. Dec. Making, 26: 285–294 (2013).
Published online 19 April 2012 in Wiley Online Library (wileyonlinelibrary.com) DOI: 10.1002/bdm.1759

Dym, C.L., Agogino, A.M., Eris, O., Frey, D.D. and Leifer, L.J. (2005), Engineering Design Thinking, Teaching, and Learning. Journal of Engineering Education, 94: 103-120. https://doi.org/10.1002/j.2168-9830.2005.tb00832.x

Engel, Jesse et al. "Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders." *Proceedings of the 34th International Conference on Machine Learning*, in PMLR. 70:1068-1077. 2017.
arXiv:1704.01279 **[cs.LG]**

Fabien, Mael. "Introduction to Automatic Speech Recognition (ASR)". *Maelfabien.github.io*. 2021. https://maelfabien.github.io/machinelearning/speech_reco/

Frayling, Christopher. “Research in Art and Design”. *Royal College of Art Research Papers*. Volume 1 Number 1 1993/4. Royal College of Art: London, 1993. https://researchonline.rca.ac.uk/384/3/frayling_research_in_art_and_design_1993.pdf

Ito, Keith, and Linda Johnson. “The LJ Speech Dataset.” Web 2017. https://keithito.com/LJ-Speech-Dataset

Kenmochi, Hideki, and Hayato Ohshita. "VOCALOID - commercial singing synthesizer based on sample concatenation". *INTERSPEECH-2007*, 4011-4010. https://www.isca-speech.org/archive/interspeech_2007/i07_4009.html

Kuligowska, Karolina et al. “Speech synthesis systems: Disadvantages and limitations”. *International Journal of Engineering and Technology(UAE)*. 7. 234-239. 2018. 10.14419/ijet.v7i2.28.12933. https://www.researchgate.net/publication/325554736_Speech_synthesis_systems_Disadvantages_and_limitations

Li, Y., Schoenfeld et al. “Design and Design Thinking in STEM Education”. *Journal for STEM Educ Res* **2,** 93–104, 2019. https://doi.org/10.1007/s41979-019-00020-z

Mulder, Hugo. “Prototyping: A Politics of Memory”. SIGRADI2018: TECHNOPOLITICAS. 22th Conference of the Iberoamerican Society of Digital Graphics. Sao Carlos: 2018. http://papers.cumincad.org/data/works/att/sigradi2018_1477.pdf

Murtaza Bulut and Shrikanth S. Narayanan. “Chapter 10 - Speech Synthesis Systems in Ambient Intelligence Environments”. *Human-Centric Interfaces for Ambient Intelligence*. Academic Press. Editor(s): Hamid Aghajan, Ramón López-Cózar Delgado, Juan Carlos Augusto. 2010, Pages 255-277. ISBN 9780123747082, https://doi.org/10.1016/B978-0-12-374708-2.00010-3. (https://www.sciencedirect.com/science/article/pii/B9780123747082000103)

Oloko-oba, Mustapha and T.S. Ibiyemi, Samuel, Osagie. (2016). “Text-to-Speech Synthesis Using Concatenative Approach”. *International Journal of Trend in Research and Development.* 3. 559-462.

Seabrook, John. *The Song Machine: Inside the Hit Factory*. W.W. Norton Co. 2016.

Shen, Jonathan et al. “Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions”. Cornell University. 2018
arXiv:1712.05884 **[cs.CL]**

Silverman, Barry G. "Expert intuition and ill-structured problem solving." *IEEE Transactions on Engineering Management,* vol. EM-32, no. 1, pp. 29-33, Feb. 1985, doi: 10.1109/TEM.1985.6447634.

Tatham, Mark and Katherine Morton. *Speech Production and Perception*. Palgrave MacMillan. UK: 2006

Van den Oord et al. “WaveNet: A generative model for raw audio”. Cornell University 2016. *https://Drive.google.com/File/D/ 0B3cxcnOkPx9AeWpLVXhkTDJINDQ/View,* arXiv:1609.03499 **[cs.SD]**

‘Voice Cloning as a Global New Technology and its Challenges for EU and Polish Law’. *KGLegal \Info Blog, IT, New Technologies, Media and Communication Technology Law.* Warsaw.
https://www.kg-legal.eu/info/it-new-technologies-media-and-communication-technology-law/voice-cloning-as-a-global-new-technology-and-its-challenges-for-eu-and-polish-law/

“Who Owns the Copyrights for Synthesized Speech”. *TTSreader*. Disqus. May 10, 2017.
https://ttsreader.com/blog/2017/05/10/copyright/

Wilkins, Julia, et al. “VocalSet: A Singing Voice Dataset.” *19th International Society for Music Information Retrieval Conference*. Paris, 2018.
http://ismir2018.ircam.fr/doc/pdfs/114_Paper.pdf

Appendix: Legal Issues

Creating a synthetic vocalist using voice input from a live performer requires an examination of what legal issues might arise. I reviewed a number of online papers to identify what legal issues might arise and strategies to address and manage those issues. The key issues identified related to copyright ownership and protection, the voice as ‘a personal right’ and privacy.

In terms of the question of copyright ownership related to outputted sound files from a speech API, the advice of a copyright attorney on the question was essentially that as long as the text used is original or otherwise legally used, any sound recordings would be copyrightable subject to the terms of use provided by the API. It was noted that copyright does not flow from the automatic process itself but rather from the creative use of the automatic process (Copyright…2017). Questions around intellectual property of the voice used in the training database are minimized since the voice is that of one singer who is a consenting participant to the use and will retain a copyright

A number of governments, including the Canadian government, have begun to identify and address digital privacy issues. The Canadian government’s response referred to as the Digital Charter Implementation Act, identifies consent as a key issue and that the consent to the provision and use of personal information be informed, freely provided and able to be withdrawn. A distinction is made between information that can be linked to an individual and information that is ‘de-identified’. A question therefore to be paid attention to and addressed if necessary, is whether any personal information can be deduced from a voice synthesizer and ensure that there is either consent and/or that the information is de-identified.

An emerging legal issue related to voice cloning technology considers the voice as ‘a personal right’ and whether the law has adequate protections. An article by a European law firm with a focus on new technologies notes that from a legal viewpoint, the voice should be ‘classified as a personal right’ and the individual’s property (KG Legal).

Copyright and privacy protection laws are potential avenues for redress in the case of a perceived violation. Though more specific legal provisions related to clarifying the voice as a personal right need to be developed, consent and clear authorization by the individual for specific uses are important to addressing potential liability issues.

As my project requires the recording of a live performer's voice to create a training database, it will be important to have the clear consent and authorization of the individual for the use of her voice. This consent should also address the issue of any privacy concerns as well as intellectual property concerns.