Proceedings of CONF-SEML 2025 Symposium: Machine Learning Theory and Applications
DOI: 10.54254/2755-2721/2025.TJ24668

Realization of Human Voice Synthesis Based on Nyquist

Cyra Chung

Santa Margarita Catholic High School, Rancho, USA
chung.cyrus@outlook.com

Abstract. As the field of computer music advances, vocal synthesis has become a prominent frontier of development. Most synthesized results rely on recorded samples as a foundation, either through concatenative synthesis or through training on large voice datasets. On this basis, this research aims to develop a voice synthesizer that does not directly depend on recorded samples, instead using only waveform synthesis. The synthesizer would produce a waveform similar to that of a human larynx, and the result would be passed through several formant filters that make up the oral and glottal cavity. Additional filters and masks would be applied to shape the sounds and produce consonants. The results were semi-successful as the synthesizer was able to produce natural sounding vowels and be modified to have different perceived ages and genders but was unable to form distinguishable consonants. Overall, these results show promise of the possibility of a natural vocal synthesizer without the usage of recorded samples, and further research would be conducted to advance its progress.

Keywords: Computer music, vocal synthesis, human voice synthesis, Vocaloid.

1. Introduction

Despite its relatively recent establishment, the field of computer music has a rich history. Developments in computer music began almost immediately after the creation of general-purpose electronic computers. As Alan Turing worked with the Manchester Electronic Computer, he recorded in his handbook the earliest known tutorial on programming a computer to produce music, using the computer's internal clock and the loudspeaker [1]. After studying Turing's notes, Christopher Strachey was able to program the British national anthem onto the computer, marking the first conventional piece of music performed by a computer. The first widely recognized computer music program, MUSIC I, was developed by pioneers, e.g., Max Mathews at Bell labs in 1957 [2]. This marked the shift from analog synthesizers toward digital sound generation, laying the groundwork for future advancements. Eventually, Mathews' innovation culminated in MUSIC V, a modular system that became foundational to later systems [3].

As computer technology developed in the 1970s and 1980s, so did computer music. Developments were made in both hardware and software, allowing for increasingly complex audio operations to be performed. Institutions like IRCAM and Stanford's CCRMA applied digital signal processing to musical contexts, refining granular synthesis, physical modeling, and spectral techniques [4]. By the 1990s and 2000s, computer music technology had become advanced enough

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.TJ24668

to be incorporated in the production of everyday contemporary music, whether it be in studios or live performances [5].

As computer music advances, vocal synthesis has become a prominent frontier of development. Vocal Synthesis is particularly intriguing as it combines the familiarity of the human voice with the artificial nature of computer technology. However, this same familiarity makes it especially difficult to create natural and realistic samples. Early attempts at voice synthesis such as DECtalk designed by Dennis Klatt were mainly based on phonetic research, focusing on intelligibility rather than realism [6]. These tools set a precedent, but they were limited in their ability to replicate the nuances of human singing. The development of concatenative synthesis was a major change in direction for vocal synthesis. By segmenting real voice recordings into individual phonemes and recombining them, systems could produce speech and melodies that sound more natural compared to early results. [7] This method was adopted by Yamaha and the Music Technology Group at Pompeu Fabra University in their creation of Vocaloid, a software that could sing user-created lyrics and melodies using pre-recorded voice banks [8]. Since its release, Vocaloid has continued to evolve, with each version producing a more natural and expressive singing voice than before [9]. More recent advancements in AI technology including deep learning and neural networks have allowed voice synthesis to become more realistic than ever before. Models such as Google's Tacotron and DeepMind's WaveNet use neural architectures to generate speech based on huge datasets [10, 11]. With slight architectural modifications and new datasets, these models would be able to generate singing as well.

This paper aims to develop a vocal synthesis program that is able to generate any voice without the need for voice actors or extremely large datasets. This program would be based almost entirely on waveform synthesis but optimized for it to sound natural. This could benefit users as they would have full control over every aspect of their vocal synthesizer and can modify it as they wish. The lack of a need to train on huge datasets is also a bonus as it saves on computing power and natural resources, making it the most environmentally friendly solution while still being versatile. The rest of the paper will detail the data required for the synthesizer, the process of synthesis, and evaluations and discussions of the results.

2. Data and method

The synthesizer will take several parameters. The first and arguably most important parameter is the fundamental frequency, which is the frequency at which the base wave will oscillate. It determines the perceived pitch of the sound and will be the main way to create melodies. The next parameter is the duration of the sound in seconds. It can be used to create short bursts of sound or sustained tones. A pair of parameters control the amplitudes of the voice and noise components of the sound. The voice component is the component generated from the base wave and the noise component is simply a band-filtered white noise. The base wave passes through a set of three formant filters, whose frequencies are controlled by three parameters. The formant frequencies determine the perceived vowel sound of the voice. Table 1 displays the frequency values of the formants for different vowels. A special parameter called the formant multiplier is used to multiply the values of all three formants. This does not change the perceived vowel but instead alters the perceived gender and age of the voice. The frequency center of the noise component is also controlled by a parameter. Different frequency centers can simulate different noise sources, such as air turbulence and fricatives. A few parameters are used to increase the realism of the voice, including the vibrato rate, tremolo rate, and the formant shifting rate. They control the rate at which the fundamental frequency of the voice, the amplitude, and the formant frequencies shift, respectively.

DOI: 10.54254/2755-2721/2025.TJ24668

The synthesizer is built from hzoscs, basic oscillators that produces sine waves at a certain frequency. They are added together to create the base wave. The noise component utilizes noise, which intuitively produces white noise. The creation of distinguishable vowels is made possible through the reason function, which acts like a band filter, focusing the base wave into formants and reducing the harshness of the white noise. Finally, lfos are used for the vibrato, tremolo, and the formant shifting, which all add to the realism of the synthesized voice.

The synthesized results will be evaluated based on whether it successfully imitates human voice and whether the vowels and consonants are understandable and distinguishable. To determine whether the synthesized voice is successful at imitating human voice, one will pass it through the track separation algorithm that are included in many DAWs, and it will be considered successful if the results are sorted under the vocal track. To determine whether the vowels and consonants are understandable and distinguishable, one will pass it through speech-to-text programs and observe whether the desired text can be produced with sufficient probability.

Table 1: Formant frequencies for five Japanese vowels

Vowel (pronunciation)	First formant (Hz)	Second formant (Hz)	Third formant (Hz)
“a” (ah)	710	1100	2540
“i” (ee)	280	2250	2890
“u” (oo)	310	870	2250
“e” (eh)	550	1770	2490
“o” (oh)	590	880	2540

3. Results and discussion

To synthesize the human voice, one had to first understand how it is created. The very first step is to reference the built-in voice extension in Nyquist, from which one learns that the voice consists of a basic wave mixed with some noise and formant filters. The basic wave determines the pitch and quality of the voice while the formant filters determine the perceived vowel and the perceived gender and age of the voice. One conducted further research on the formation of human voice, and one learned that the basic wave should be similar to a sawtooth wave due to the way the glottis vibrates. One also learned that there should be comb-filtered noise around the harmonics of the saw wave because of the imperfect nature of the human body. Additionally, one learned that the human voice is shaped by various parts of the vocal tract, which have different resonating frequencies, creating what is known as formants, which are frequency bands that are amplified. One obtained the values of the formant frequencies for different vowels of an average adult male, which will be stored as constants in the program.

For the Nyquist program, this paper begins with creating the base wave. This research will start building a saw wave by adding sine waves with amplitude 1/n at the nth harmonic. One stops at 30 harmonics because of two reasons: first, the additional harmonics do not add much content to the sound as their amplitudes are very low; second, a frequency greater than half of the sampling rate can cause aliasing, which creates artifacts and distorts the sound. One can attribute the lack of extremely high frequencies to the imperfections of the human body. The resulting waveform currently sounds too harmonic and unnatural for a human voice, so one needs to add noise in order to make the waveform sound more natural and realistic. Adding plain white noise does not create a desirable result, as the saw wave and the noise does not blend well. Adding comb-filtered noise with

DOI: 10.54254/2755-2721/2025.TJ24668

peaks that are aligned with the harmonics of the saw wave does seem to work, as it results in a waveform with more texture but still is clearly a saw wave. The strength and width of the noise can be tweaked to modify the texture of the waveform.

Then, this study moves on to adding the formant filters. One can implement formant filtering with three formants by simply passing three copies of the base wave into resonance or band-pass filters that have frequency centers at the three formants. Currently, the formants are passed directly as parameters, but they could be stored as a table of constants for ease of access. Interestingly, multiplying all the formant frequencies by a constant value does not change the perceived vowel, but it alters the perceived gender and age of the voice. This is because growth and especially androgenic puberty cause each part of the vocal tract to lengthen, thus lowering their resonant frequencies. By studying the human voice in a spectrogram, one can roughly estimate the width of each formant. One settled on 50 Hz, 100 Hz, and 200 Hz, for the first, second, and third formant, respectively. One also noticed that the first formant is usually the strongest, having the greatest amplitude, while the second and third formants are weaker in comparison. One accounts for this by multiplying the waves of the second and third formants by a constant coefficient of 0.5 and 0.25, respectively. It is possible and very likely that these constants vary from person to person, or even from different sounds produced by one person, so one could change them to be parameters so that users could precisely tweak them to produce their desired result. However, for the purposes of this project, changing these parameters does not have a significant effect on the quality of the synthesized voice, which is why one has left them as constants.

One can further refine the synthesized voice and make it more realistic. The most effective way is to add variance to the sound, as a completely static voice sounds robotic and unnatural. The first place to add variance to is the base wave, where one adds a slight vibrato, as the pitch of the human voice is generally constantly shifting, and it would be very difficult to maintain a perfectly constant pitch. Next, this research will slowly shift the formant centers because it is again very difficult to maintain a constant resonant frequency in one's vocal tract. Then, one adds amplitude modulation to the sound, which greatly increases the realism of the voice as the amplitude of the human voice is always varied. Additionally, one can add a separate noise component to the voice at a higher frequency to mimic the breathiness of the voice. This noise component could also be useful for synthesizing consonants, as most involve creation of noise as air passes through a more restrictive vocal tract. Finally, one passes the voice through an envelope to obtain a more natural attack and release. This gives us the final result: a realistic synthesized voice capable of singing different pitches and vowels and having different perceived age and gender.

The results of the voice synthesizer certainly resemble human voice with clearly distinguishable vowels. Produced samples consistently pass the track separation test, being sorted under the vocal track with perfect accuracy. The vowels can be understood by speech-to-text programs, usually performing better when multiple samples are played back-to-back. This could be because multiple samples allow speech-to-text programs to better perceive the formant differences in the sound.

Limitations and Prospects

The most significant limitation of the synthesizer is its current inability to produce distinguishable consonants. Consonants are a crucial part of language and words cannot be formed without them. The difficulty of producing consonants lies in the fact that they consist of mainly noise, which is harder to distinguish than periodic waveforms. After the production of consonants, the next goal would be to implement a function that takes a list of parameters, including the pitches, durations, and lyrics, and returns an entire phrase. As of now the implementation of the synthesizer is very limited, only being able to produce a single constant tone at a time and having to manually

DOI: 10.54254/2755-2721/2025.TJ24668

pass the three formants for the desired vowel. While piecewise functions can be used to create changing sounds in a single breath, it is very awkward to use, and the synthesizer would benefit greatly from better implementation. Another step would be to make the synthesizer more customizable, allowing the user to tweak every single parameter to their liking, such as the widths of the formant filters, in order to create their perfect Vocaloid. Finally, the project can be finished with expansion beyond the Japanese language by being able to take each sound from the international phonetic alphabet separately, gaining access to unrestricted phonotactics and allowing the synthesized voice to possibly sing in every language in the world.

4. Conclusion

To sum up, vocal synthesis, just like any other field of study, will continue developing as technology advances. However, as technologies such as AI advance, one has to be conscientious of the impacts of large-scale training in data centers on the environment and the world. The results have shown that it would be possible to synthesize natural and realistic vocals without the need for extensive training on large datasets, and that further research is needed to turn it into a reality. Ultimately, as one enjoys the comfort and entertainment that technology brings to mankind, one must not forget to use the own intelligence to protect the natural world as well.

References

[1] Copeland, B.J. and Long, J. (2017) Turing and the History of Computer Music. Boston Studies in the Philosophy and History of Science, 189–218.
[2] Chadabe, J. (1997) Electric Sound: The Past and Promise of Electronic Music. Prentice Hall.
[3] Roads, C. (1996) The computer music tutorial. MIT Press.
[4] Wishart, T. (1996) On sonic art (2nd ed.). Routledge.
[5] Emmerson, S. (2007) Living Electronic Music. Ashgate Publishing.
[6] Klatt, D.H. (1987) Review of text-to-speech conversion for English. The Journal of the Acoustical Society of America, 82(3), 737–793.
[7] Stylianou, Y. (2001) Applying the harmonic plus noise model in concatenative speech synthesis. IEEE Transactions on Speech and Audio Processing, 9(1), 21–29.
[8] Kenmochi, H. and Ohshita, H. (2007) VOCALOID—Commercial singing synthesizer based on sample concatenation. Proceedings of Interspeech 2007, 4011–4010.
[9] Advantages of VOCALOID6 - VOCALOID - the modern singing synthesizer -. (2022, August 16). Retrieved from https: //www.vocaloid.com. https: //www.vocaloid.com/en/vocaloid6/
[10] Shen, J., Pang, R., Weiss, R.J., et al. (2018) Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. ICASSP 2018 - IEEE International Conference on Acoustics, Speech and Signal Processing, 4779–4783.
[11] van den Oord, A., Dieleman, S., Zen, H., et al., (2016) WaveNet: A generative model for raw audio. arXiv preprint arXiv: 1609.03499.