University of Waterloo

CS 489 Computational Audio

Investigating the Fidelity of Generative AI in Voice Recreation

March 15, 2024

Sahl Bakshi

Topic

The primary objective of this project is to assess the capability of generative AI voices to faithfully recreate the unique characteristics of real human voices. Through the comparison of human and AI generated voice samples, I aim to investigate the potential for using AI to perfectly replicate human speech.

Data Collection

Human Voice Samples

I carefully selected and recorded one of many spoken paragraphs in a quiet room, ensuring minimal background noise to create a high-quality dataset for training the model. With this dataset, I anticipate that the voice clone will accurately replicate my speech patterns and vocal characteristics, delivering a seamless and authentic voice experience. I used the following paragraph as my dataset.

“Cathy is sixteen years old. She's excited to learn to drive. Her mom, who has a driver's license, is her instructor. They practice driving twice a week. Cathy has a learner's permit, allowing her to drive with an adult. Today, they're tackling the freeway for the first time. Both Cathy and her mom are nervous. The freeway is crowded with cars and trucks. Cathy doesn't like driving fast or being near big trucks. As they merge onto the freeway, Cathy's grip on the wheel tightens. She's anxious but determined. With her mom's encouragement, Cathy starts driving slowly. She feels the car's engine hum beneath her. Despite her nerves, Cathy steers carefully through traffic. The speed increases, making Cathy's heart race. She takes deep breaths to calm herself. Cathy focuses on the road ahead, her confidence growing with each passing mile. Finally, she realizes she's successfully driving on the freeway!

As they continue along the freeway, Cathy's mom provides gentle guidance, helping her navigate the lanes and merge smoothly. Despite the initial anxiety, Cathy begins to feel more at ease behind the wheel. The scenery whizzes by, and Cathy marvels at how far she's come since the beginning of her driving lessons. With each passing mile marker, Cathy's confidence blossoms, and she starts to enjoy the sensation of driving on the open road. She glances at her mom, who smiles proudly, reinforcing Cathy's sense of accomplishment and reassuring her that she's doing great. Cathy's nerves gradually dissipate, replaced by a newfound sense of freedom and independence as she continues her journey down the freeway.”

Generative Voice Samples

Initially, my intention was to utilize the voice cloning feature offered by https://docs.app.resemble.ai/ to replicate my voice using python and the above dataset I had prepared. However, upon considering the cost involved with this tool, I opted for an alternative solution provided by https://play.ht/voice-cloning/. This platform offers an integrated tool for

generating voices by uploading a WAV, MP3, or any other sound file and is trusted my many reputable organizations. For optimal results, I ensured that my recording was a minute long to facilitate better cloning. When describing my voice, I simply stated, "a male university student aged 22", providing minimal information. Additionally, I had the option to choose between an enhanced version (cloned from a de-noised, re-mastered version of the original audio) and normal version (cloned from the original audio without any changes), ultimately selecting the enhanced version for the best possible outcome.

Analysis

Waveform Analysis

I used audacity to get the waveforms of both audio files as shown.

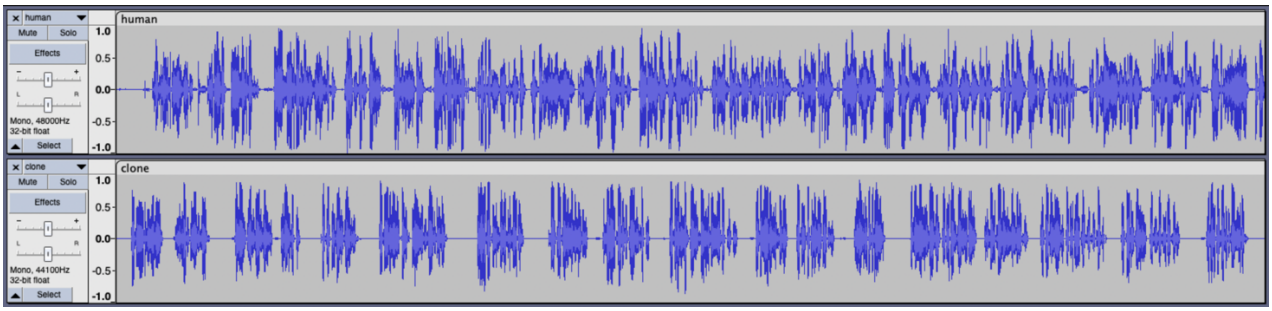


Figure 1 Waveforms of human and cloned voice

Amplitude and Loudness: Both waveforms exhibit varying heights, which represent the amplitude of the sound wave. The amplitude correlates with volume (loudness), and in both waveforms, we see similar patterns of variation. This suggests that both the human and cloned voice have a similar dynamic range.

Waveform Shape and Timbre: The shape of the waveforms relates to the timbre or quality of the sound. Although visually it's difficult to deduce precise timbral qualities, the fact that both waveforms have a similar overall shape indicates that the tonal qualities of the human voice might have been reasonably captured by the cloned voice.

Frequency and Pitch: Frequency, which is related to pitch, isn't directly visible in these waveforms since it would require a spectrogram or a frequency analysis. However, the spacing between peaks can give a crude idea of frequency content. The spacing seems quite similar in both, suggesting that the pitch may be closely matched.

Temporal Features: When looking at a waveform, the temporal features such as attack, decay, sustain, and release of sound can sometimes be inferred. Both waveforms show similar

temporal patterns, which suggests that the cloned voice follows the same speech cadence and rhythm as the human voice.

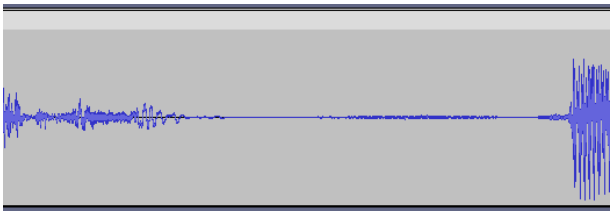
Sampling Rate: Notably, the "human" waveform is labeled as having a sampling rate of 48000 Hz and the "clone" waveform has a sampling rate of 44100 Hz. This difference in sampling rates usually affects the frequency content that can be accurately represented. The Nyquist Theorem states that the sampling frequency must be at least twice the highest frequency contained in the signal to be sampled without aliasing. In practice, this difference should not have a significant impact on human voice perception since both exceed the typical maximum frequency of human speech (~20 kHz at most).

Bit Depth: Both waveforms are labeled as 32-bit float, which means that they have a high dynamic range and can capture both very soft and very loud sounds with precision. This is a good choice for professional audio work as it allows for more headroom in processing and editing.

Finally, from a visual standpoint, the two waveforms exhibit strong similarities, which implies that the cloned voice is a close approximation of the human voice in terms of volume, temporal characteristics, and possibly pitch. However, an accurate technical analysis would require listening to the actual audio and using spectral analyzers both of which is down later.

Zoomed Comparison

As suggested, I focused on zooming in the two voice samples at different parts of the audio clips. Here is what I found.



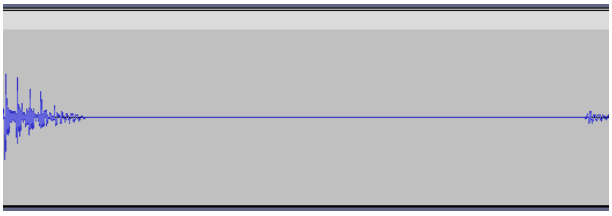


Figure 2 & 3 Zoomed in quiet parts of human and cloned speech respectively

The second waveform is cleaner in between louder audio, with a sharp spike at the beginning and a clear end. The uniformity and lack of irregularities could suggest that this is a digitally generated or processed sound, aligning with the "clone" voice. The clean stops could be a result of digital processing where the voice synthesis is engineered to have

precise beginnings and endings without the natural decay or ambient noise typically found in human recordings.

The first waveform is more complex with several fluctuations in amplitude throughout the recording. The irregularities at the ends, described as "small disturbances of stops," are common in natural speech. These could be the result of the speaker taking breaths, movements, or subtle shifts in the environment that are captured by the microphone. Such nuances are often smoothed out or entirely absent in the synthetic voice.

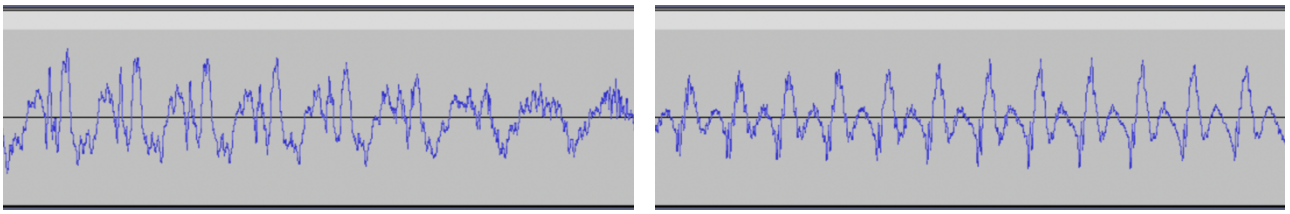


Figure 4 & 5 Zoomed in talking parts of human and cloned speech respectively

Clone waveform is very uniform and regular. The peaks and troughs are consistent, which may indicate a synthetic or processed origin. The amplitude changes are very structured, possibly suggesting the use of a voice synthesis algorithm that creates smooth transitions. The absence of erratic or abrupt changes suggests less variation in vocal intensity or emotion.

Human waveform shows more variation in amplitude and frequency. The irregularity of the peaks and troughs reflects the natural variation in human speech. The less predictable pattern is indicative of the natural fluctuations in pitch and tone that occur as a person speaks. There may be evidence of breaths, mouth movements, or other non-verbal sounds that humans naturally produce while talking.

Therefore, when zoomed in we can see that the human voice waveform shows a more complex and variable structure, indicative of the natural inconsistencies and inflections that occur during speech while the cloned voice waveform has a more mechanical and repetitive structure and lacks the subtleties and expressiveness found in the human voice.

Stretched Comparison

I stretched both the clips by going to Audacity → Effect → Pitch and Tempo → Change Tempo and 4xed the time for both clips as shown below.

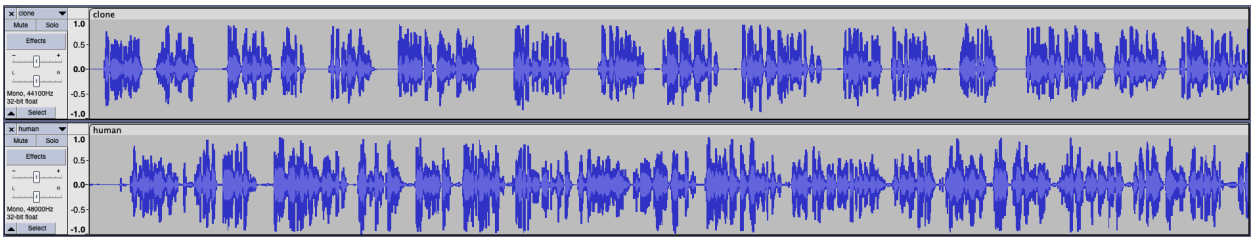


Figure 6 Waveforms of human and cloned voice stretched

What we found even after zooming in and inspecting both waveforms is the same conclusion that we find when inspecting them regular without changing the tempo. This process usually spreads out the waveform, making each individual cycle of the wave more visible and easier to analyze. However, even after changing the tempo and zooming in on the waveforms, it did not make it easier to distinguish between the two.

Spectrogram Analysis

I used audacity to get the spectrograms of both audio files as shown.

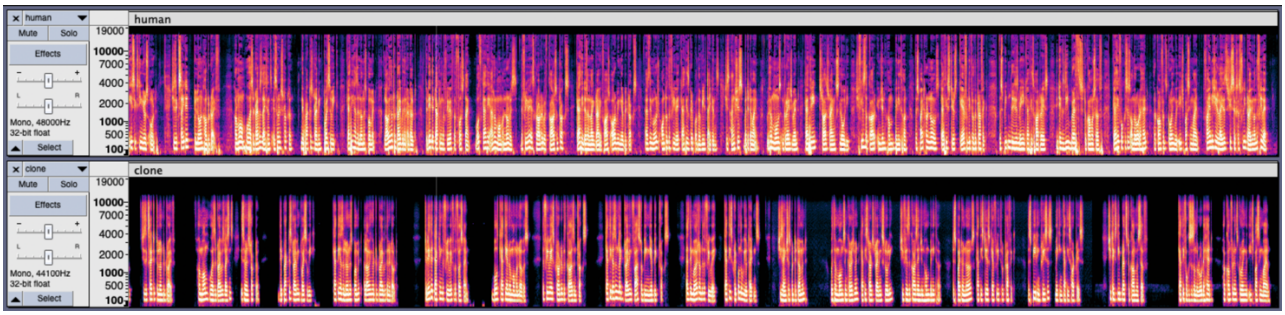


Figure 7 Spectrograms of human and cloned voice

Frequency Information: The vertical axis represents frequency, with lower frequencies at the bottom and higher frequencies at the top. Both spectrograms cover the same frequency range, which is typical for human speech, extending up to 19,000 Hz. This suggests that both the human and cloned voices have been analyzed over the same frequency spectrum.

Time Evolution: The horizontal axis represents time. The patterns over time appear to be very similar, with the "events" in the human voice finding corresponding events in the cloned voice. This implies that the cloned voice has a similar rhythm and timing to the human voice.

Intensity and Color Coding: The colors in a spectrogram usually indicate the intensity (or power) of frequencies at each point in time, with brighter colors representing higher intensity. Both spectrograms seem to use similar color schemes, and there is a close match in the distribution of colors, which suggests that the volume and emphasis of certain frequencies are being replicated by the cloned voice.

Harmonic Structures: The series of horizontal lines seen in both spectrograms represent the harmonic content of the voice. These are multiples of the fundamental frequency (the lowest frequency of a sound). The human voice has a clear set of harmonics, and it appears that the cloned voice reproduces these with some degree of accuracy, indicating a similar pitch and timbral quality.

Formants: Formants are the concentration of acoustic energy around certain frequencies and are crucial in characterizing vowel sounds. In both spectrograms, there are similar patterns that indicate the formants are being closely matched by the cloned voice. This is important for ensuring that vowels sound the same in both the human and cloned recordings.

Sampling Rate: The "human" spectrogram is from a recording with a 48000 Hz sampling rate, while the "clone" is from a 44100 Hz sampling rate. This slight difference shouldn't greatly affect the visible spectrogram for voice recordings, as mentioned previously, because the frequency content of a human voice typically falls well below the Nyquist frequency of both sampling rates.

Differences: While there is a strong resemblance, some subtle differences may be present, possibly due to the inherent variability in producing the same vocal sound or limitations in the cloning technology. These might show up as slight discrepancies in the intensity and exact position of the harmonics and formants.

In conclusion, the cloned voice's spectrogram shows a high degree of similarity to the human voice's spectrogram, suggesting that the clone captures the fundamental frequency, harmonics, formants, and temporal patterns of the human voice with reasonable accuracy. It is worth noting, however, that visual inspection of spectrograms can only reveal so much; an auditory analysis would be required to determine the perceptual similarity between the two voices which is down below.

EAC Option

The previous analysis had the default Algorithm option set to Frequencies with the Scale option set to MEL. As suggested, I compared the speeches with the Algorithm set to Pitch (EAC) and the Scale still set to MEL spacing as shown.

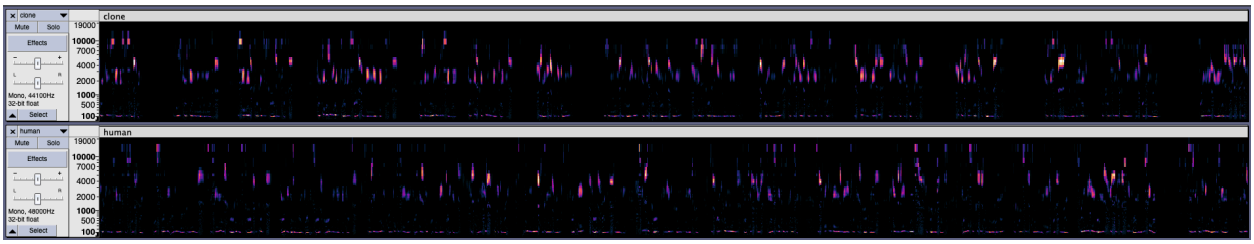


Figure 8 Spectrograms of human and cloned voice with above configurations

Overall Pattern: Both spectrograms show a pattern of activity that suggests speech, with distinct breaks and variations that are typical of spoken language.

Harmonics: There appears to be harmonic structures in both samples. Harmonics are the horizontal lines above the fundamental frequencies (the brightest lines), indicating that both voices have a richness or timbre that includes overtones.

Temporal Features: The spacing between the vertical clusters seems similar, indicating that the rhythm or timing of speech are comparable between the two samples.

Frequency Range: The "clone" voice seems to have a slightly narrower frequency range compared to the "human" voice. This can be seen in the spread of the harmonics and the overall height of the brighter areas in the spectrogram.

Brightness and Clarity: The "human" sample appears to have brighter and more defined areas of intensity, which suggests a greater dynamic range or a more pronounced articulation of certain phonetic components.

Harmonic Structure: While both show harmonics, the "human" spectrogram has more clearly defined harmonic lines. The "clone" sample's harmonics seem to be less distinct and more diffused.

Pitch Consistency: There might be slight variations in the pitch consistency of the "clone" sample when compared to the human one. This could be seen where the bright lines in the clone sample appear more wobbly or less straight than in the human sample.

Noise and Artifacts: The "clone" might show more noise or artifacts, visible as random specks of colors in the spectrogram, which could be a result of the synthesis process.

These observations suggest that while the cloning process has been able to capture the general pattern and structure of human speech, there are still noticeable differences in the quality and definition of the sounds produced. The clone's voice might lack the same richness, clarity, and variability that the human voice naturally has.

Subjective Evaluation

To assess the accuracy of my voice cloning, I conducted a series of experiments involving friends, family, and other individuals. The methodology involved presenting only my cloned voice initially, followed by the introduction of my original voice. The results of this comparative analysis are summarized below. In total, 34 participants took part in the evaluation, 10 of whom I knew prior to the assessment.

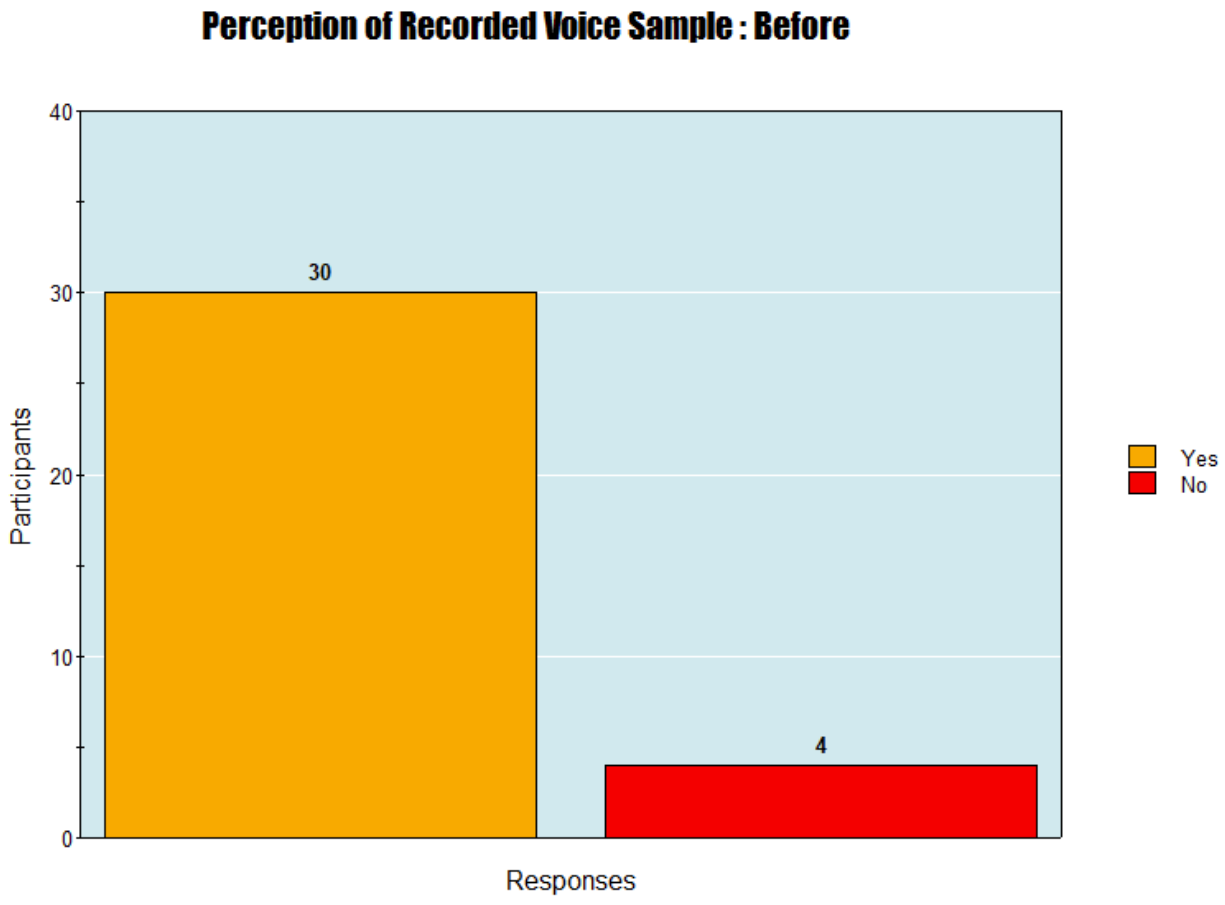


Figure 9 Perception of Recorded Voice Sample before introduction of original voice

The results indicate that most participants perceived the voice as original rather than cloned. Interestingly, many participants also speculated that the voice was recorded in a podcast or similar setting due to the absence of background noise.

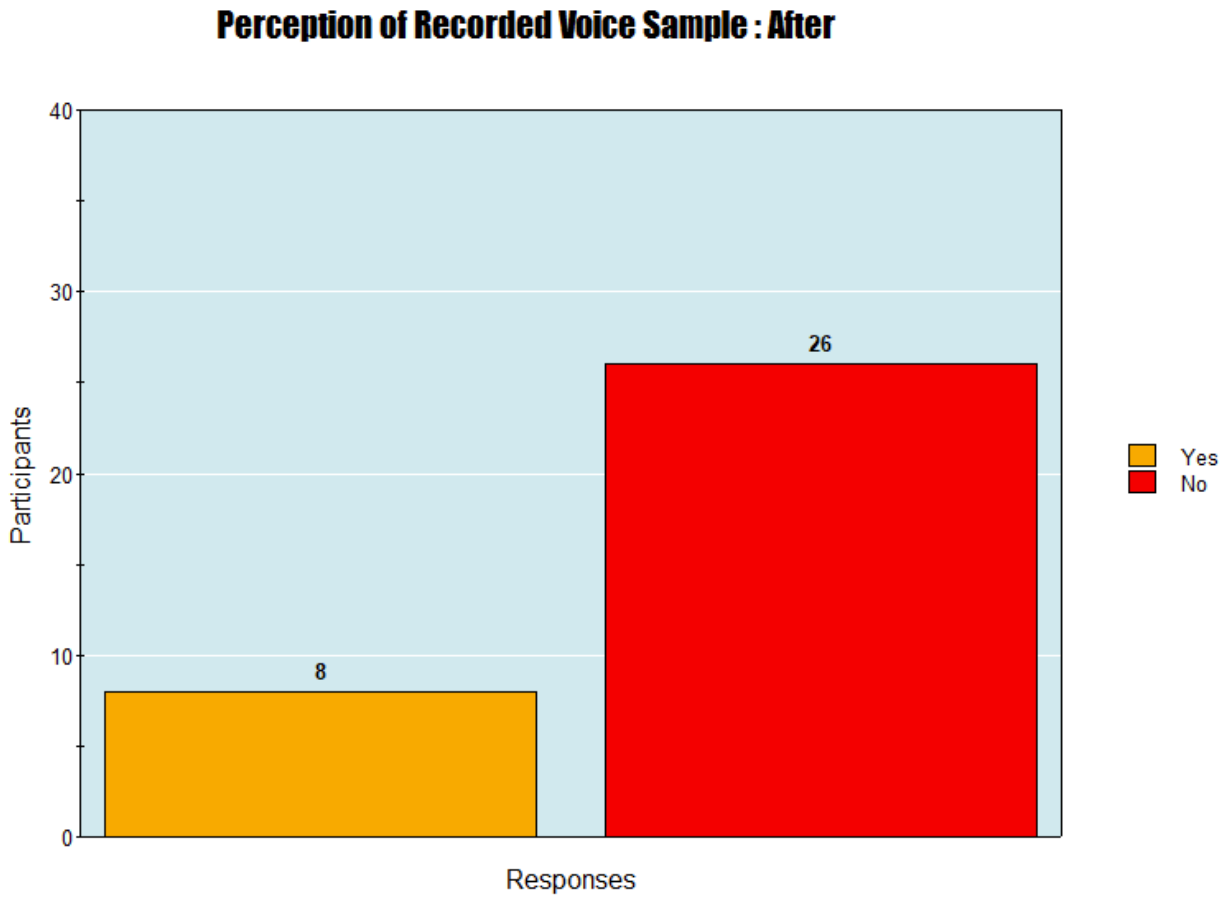


Figure 10 Perception of Recorded Voice Sample after introduction of original voice

Following these results, although some participants still believed that the cloned voice was original without noise, the majority concluded that the voice was indeed cloned. However, the primary reason for this consensus was the absence of natural speech cues such as pauses, breaths, "umms", and "hmms", leading them to believe it was generated. Despite this, the cloned voice was nearly indistinguishable from the original. These results are in line with an article from Speech Technology News [1] saying that “only one in 50 people can correctly identify artificial intelligence-generated voices compared to the real human ones,

a study by cloud communications company Ringover found”.

Conclusion

The exploration into generative AI's capacity to emulate human voice characteristics showcases a significant stride towards achieving realistic voice replication. Through the data collection involving high-quality human voice samples and the application of advanced voice cloning technologies, this project studied the nuances of speech patterns, volume, tempo, and timbre that define individual human voices. The comparison of waveform and spectrogram analyses between original and generated voice samples reveals a striking resemblance in amplitude, frequency, and temporal features, underscoring the cloned voice's fidelity to the human original. Moreover, the subjective evaluation involving participants unfamiliar with the nuances of voice cloning further illuminates the AI's proficiency in generating convincingly authentic voice replicas.

While the absence of natural speech cues in AI-generated voices was noted, the overall perception remained that the cloned voice bore a near-indistinguishable resemblance to the original. This convergence of objective and subjective assessments underscores the potential of generative AI technologies to create highly realistic voice clones, albeit with room for improvement in simulating the subtleties of human speech dynamics. This project not only demonstrates the current capabilities of AI in voice cloning but also opens avenues for future enhancements, particularly in integrating natural speech nuances to enhance realism further.