How Does the Optimal Number of Fourier Series Terms Required to Reconstruct an Alexa Voice Command Vary Between Different Speakers?

Aarush Raheja

Abstract

This study investigates whether the number of Fourier series terms required to reconstruct the same spoken Alexa command varies between speakers. Two controlled recordings, representing a softer voice and a louder, deeper voice, were modeled as finite Fourier series. Reconstruction quality was evaluated using root mean square error (RMSE) and the proportion of Fourier-domain energy captured. The optimal order was defined as the smallest number of terms that captured at least 95% of the reference energy while producing less than a 1% relative RMSE improvement when another term was added. The soft sample required 193 terms and the louder sample required 194 terms. The result indicates a small speaker-dependent difference for the tested recordings, although the narrow sample limits generalization.

Keywords: Fourier series; speech reconstruction; signal processing; RMSE; spectral energy

1. Introduction

Voice assistants interpret commands spoken by users whose voices differ in pitch, tone, articulation, and intensity. Although two recordings may contain the same words, their time-domain waveforms and frequency distributions can be noticeably different. Fourier analysis provides a mathematical framework for representing these signals as combinations of sinusoidal components.

A Fourier series represents a periodic function as an infinite sum of sine and cosine terms (Weisstein, 2008). A practical reconstruction must use only finitely many terms, creating a tradeoff between signal fidelity and computational cost. This study asks: How does the optimal number of Fourier series terms required to reconstruct an Alexa voice command vary between different speakers?

The same command was recorded under controlled conditions in a soft voice and a louder, deeper voice. The analysis compares the number of harmonics required under two stopping conditions: captured spectral energy and diminishing improvement in RMSE.

2. Mathematical Framework

2.1 Modeling a voice signal

A recorded sound wave is modeled as a real-valued amplitude function f(t) over a finite interval $-T/2 \le t \le T/2$. Digital audio stores this signal as discrete samples taken at a constant sampling rate. Because natural speech is not periodic, the selected interval is treated as one period and extended according to $f(t) = f(t + T)$. This approximation permits Fourier-series analysis while retaining the structure of the recorded window.

2.2 Fourier representation

The Fourier series of a period-T signal is

$$f(t) = a_0/2 + \Sigma_{n=1}^{\infty} [a_n \cos(2\pi nt/T) + b_n \sin(2\pi nt/T)].$$

Sine and cosine functions of different integer frequencies are orthogonal over one period. Therefore, cross-products integrate to zero, while the squared sine or cosine integrates to T/2. These relations isolate the coefficients:

$$a_0 = (2/T) \int_{-T/2}^{T/2} f(t)\, dt,$$
$$a_n = (2/T) \int_{-T/2}^{T/2} f(t) \cos(2\pi nt/T)\, dt,$$
$$b_n = (2/T) \int_{-T/2}^{T/2} f(t) \sin(2\pi nt/T)\, dt.$$

A finite reconstruction of order N retains the first N harmonics:

$$S_N(t) = a_0/2 + \Sigma_{n=1}^{N} [a_n \cos(2\pi nt/T) + b_n \sin(2\pi nt/T)].$$

The approximation error is $e_N(t) = f(t) - S_N(t)$. Lower-order terms describe broad waveform behavior; higher-order terms reproduce faster changes and finer detail.

2.3 Evaluation metrics and stopping rules

For a discrete signal x[k] sampled at M equally spaced times, RMSE was used to measure time-domain reconstruction error:

$$RMSE(N) = \sqrt{\{(1/M) \Sigma_{k=1}^{M} [x[k] - S_N(t_k)]^2\}}.$$

The relative improvement from one additional term was defined as $r(N) = [RMSE(N-1) - RMSE(N)] / RMSE(N-1)$. The RMSE condition was satisfied when $r(N) < 0.01$.

Parseval's theorem relates the energy of a signal to the squares of its Fourier coefficients (Innovation World, 2025). For the finite series, captured energy was

$$E_N = a_0^2/4 + \Sigma_{n=1}^{N} (a_n^2 + b_n^2).$$

The reference order was $N_{max} = 200$, and the captured proportion was $P(N) = E_N/E_{200}$. The optimal order N* was the smallest value satisfying both $P(N) \geq 0.95$ and $r(N) < 0.01$.

3. Method

The spoken command “Alexa, what's the weather like?” was recorded twice: once in a soft voice and once in a louder, deeper voice. The same external microphone and recording device were used. The speaker remained 40 cm from the Alexa device, and recordings were made in a quiet indoor environment. Each file was trimmed to three seconds so that silence before and after the command was excluded.

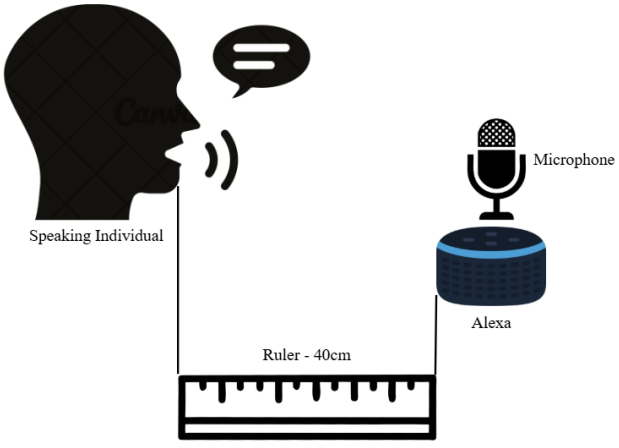

Figure 1. Controlled voice-recording setup. Source: author-designed diagram.

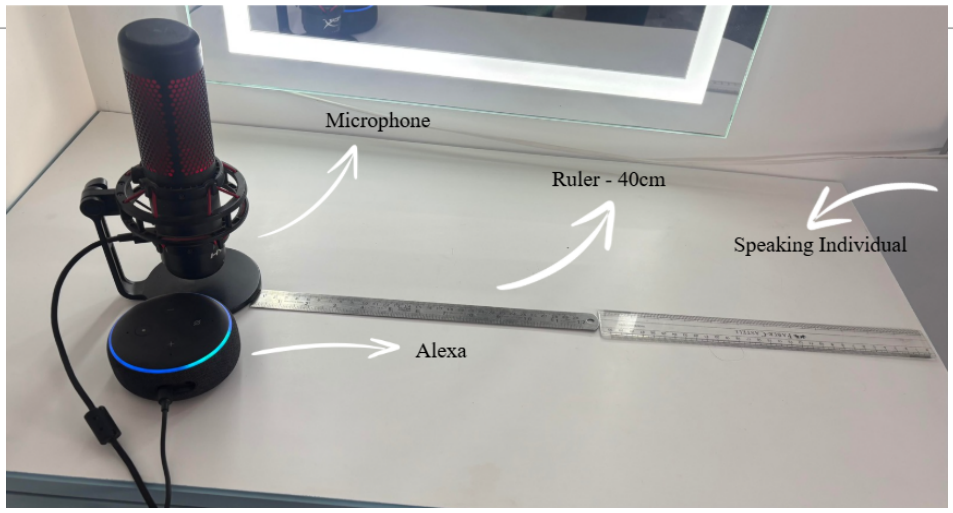

Figure 2. Experimental setup used for audio collection. Source: author photograph.

The MP3 files were imported into Python and converted to numerical arrays using librosa and SciPy. Fourier coefficients were approximated by discrete summation for increasing values of N. For each order, the reconstructed signal, RMSE, relative RMSE improvement, captured energy, and captured-energy proportion were computed. Values from 1 through 200 were tested.

4. Results

4.1 Soft voice sample

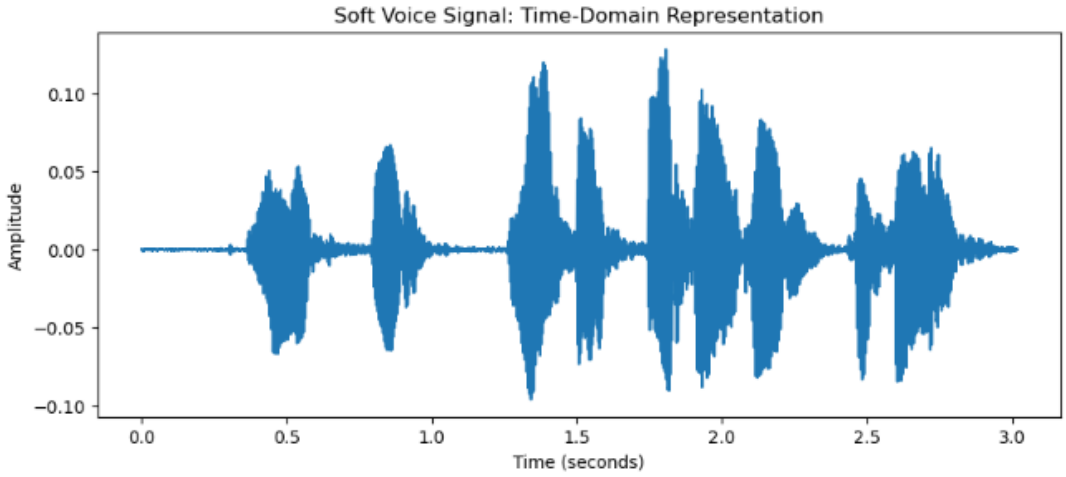

Figure 3. Time-domain waveform of the soft voice sample. Source: author computation.

At N = 1, the reconstruction was smooth and retained almost none of the reference Fourier energy. RMSE was 0.0218, E_1 was 2.72×10^{-11}, and P(1) was 3.83×10^{-5}. The first harmonic alone therefore did not represent the command's detailed structure.

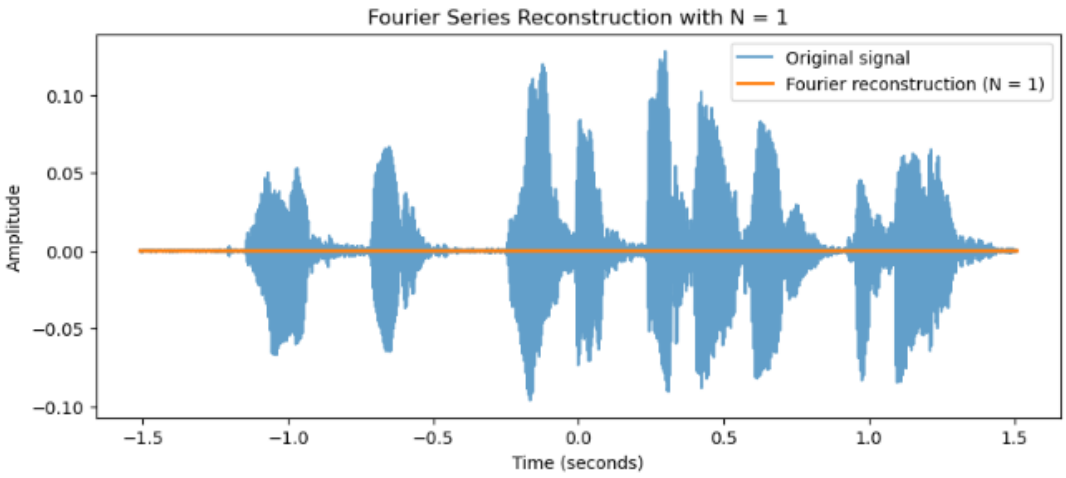

Figure 4. Soft-voice reconstruction using one Fourier term. Source: author computation.

At N = 5, RMSE was 0.02183 and the improvement ratio was 9.89×10^{-4}, satisfying the diminishing-RMSE condition. However, P(5) was only 5.43×10^{-4}, so the energy condition was not satisfied.

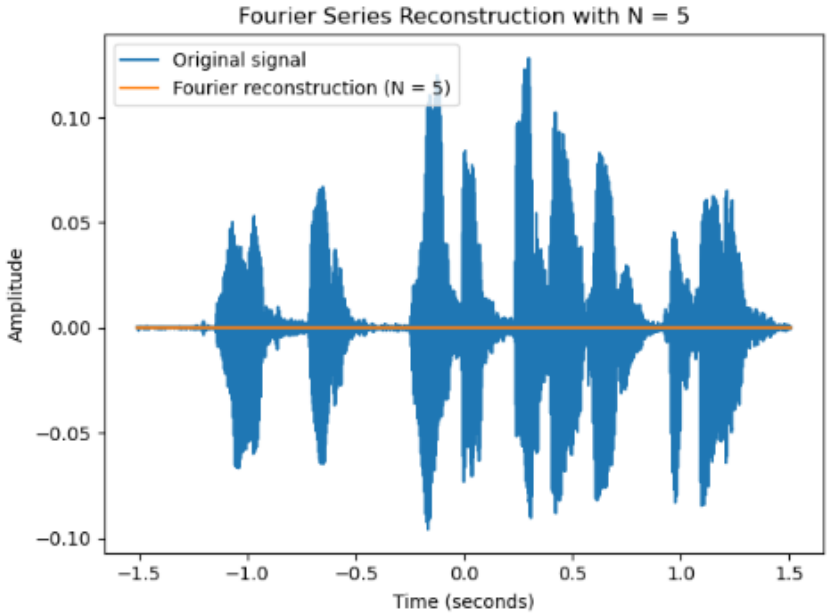

Figure 5. Soft-voice reconstruction using five Fourier terms. Source: author computation.

The first order satisfying both conditions was N = 193. At this order, RMSE was 0.02088, r(193) was 1.14×10^{-5}, E_{193} was 6.764×10^{-7}, and P(193) was 0.953. The reconstructed waveform closely followed the timing and amplitude structure of the original sample.

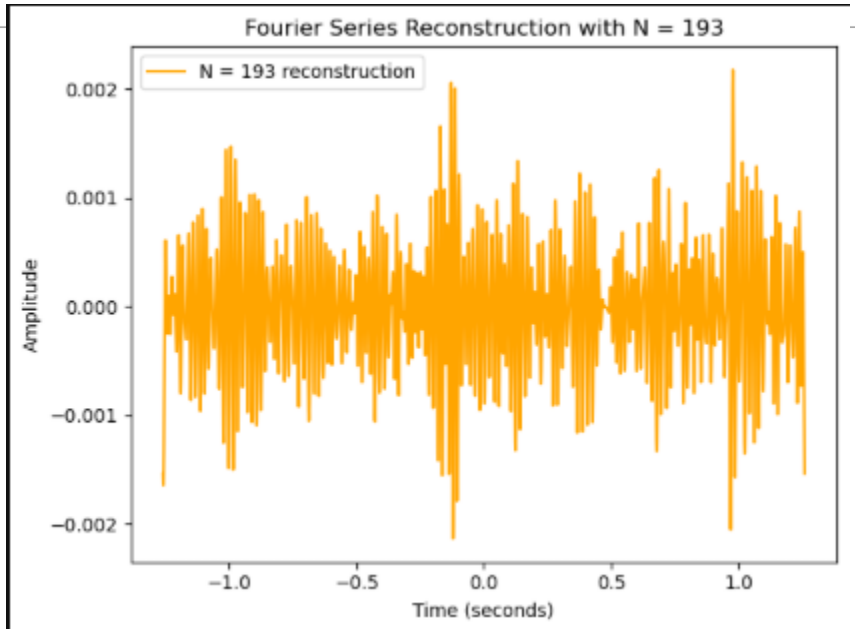

Figure 6. Soft-voice reconstruction using 193 Fourier terms. Source: author computation.

N	RMSE	r(N)	Energy E_N	P(N)
1	0.0218	-	2.72 × 10□¹¹	3.83 × 10□□
5	0.02183	9.89 × 10□□	3.90 × 10□¹□	5.43 × 10□□
193	0.02088	1.14 × 10□□	6.76 × 10□□	0.953

Table 1. Selected soft-voice reconstruction metrics.

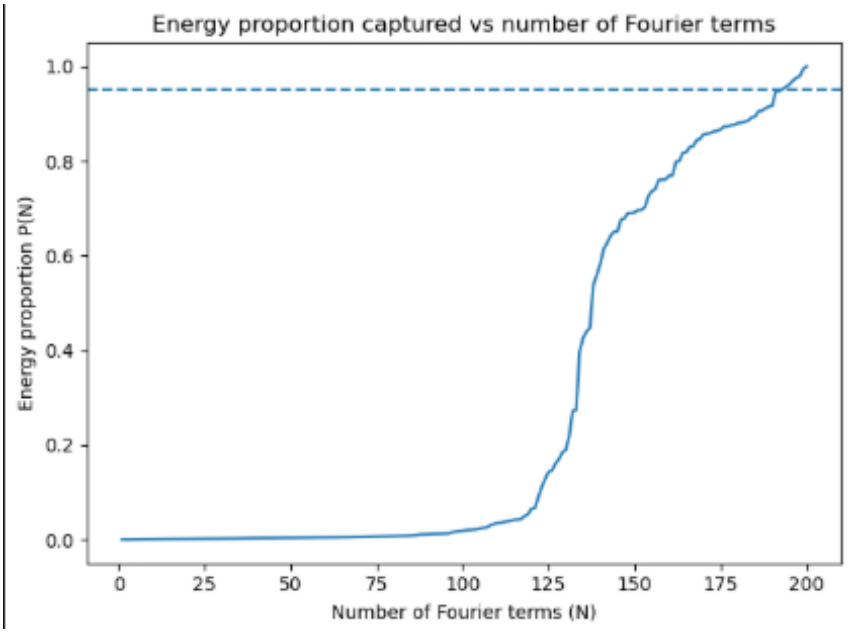

Figure 7. Captured-energy proportion as a function of Fourier order for the soft voice. Source: author computation.

4.2 Louder voice sample

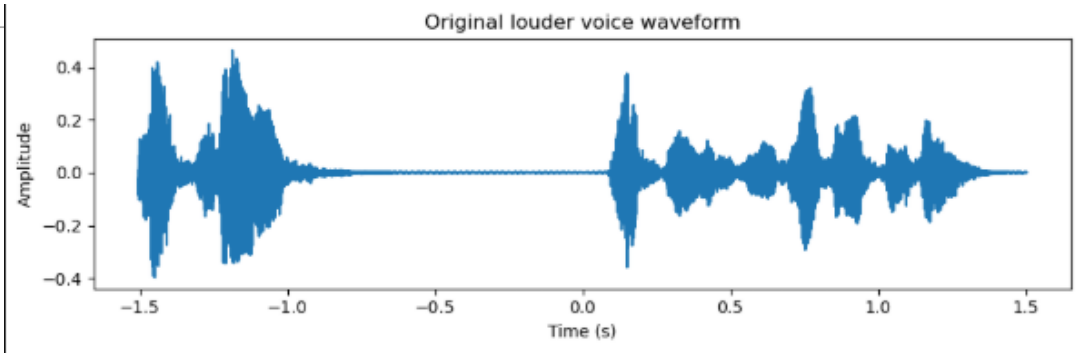

Figure 8. Time-domain waveform of the louder voice sample. Source: author computation.

The louder sample reached the joint stopping criterion at N = 194. At this order, RMSE was approximately 0.0560, r(194) was 2.60×10^{-7}, E_{194} was 1.77×10^{-6}, and P(194) was 0.957. Energy accumulated slowly at low orders and rose more rapidly among the higher harmonics.

N	RMSE	r(N)	Energy E_N	P(N)
1	0.0560	-	1.74 × 10□□	0.00075
5	0.0560	1.30 × 10□□	9.59 × 10□□	0.00413
194	0.0560	2.60 × 10□□	1.77 × 10□□	0.957

Table 2. Selected louder-voice reconstruction metrics.

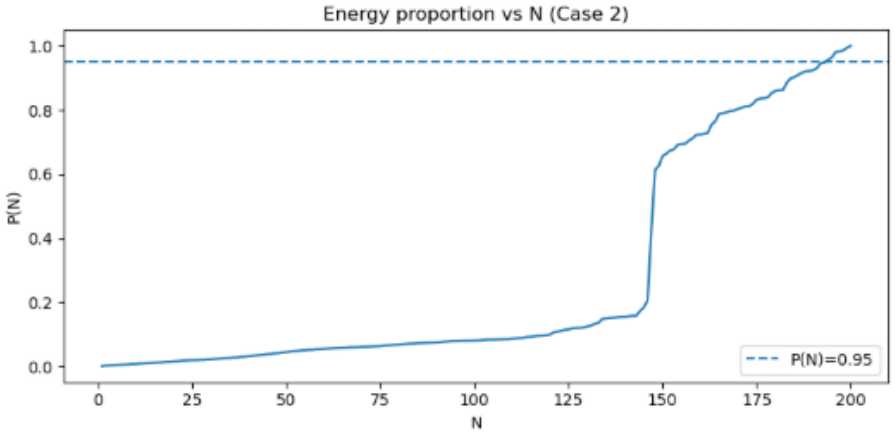

Figure 9. Captured-energy proportion as a function of Fourier order for the louder voice. Source: author computation.

5. Discussion

The soft and louder samples required 193 and 194 terms, respectively. The difference of one harmonic supports the proposition that the optimal order can vary between speakers, but the observed effect is small. The louder signal also had a larger RMSE and reference energy, consistent with greater amplitude variation and a different spectral distribution.

The result depends strongly on the operational definition of optimality. Because reference energy was normalized to N_{max} = 200, an order near 200 was likely when high-frequency components contributed substantially to the selected window. A different maximum order, energy threshold, window length, or perceptual metric could change the selected values. In addition, RMSE changed little across the tested orders, showing that time-domain error and coefficient energy measure different aspects of fit.

Only two recordings were analyzed, and they were described as voice conditions rather than a broad, demographically varied speaker sample. Consequently, the difference cannot be generalized to populations or attributed specifically to loudness, pitch, or speaker identity. The periodic-extension assumption can also introduce discontinuities at the edges of the three-second window.

Future work should include more speakers, repeated trials, calibrated sound-pressure levels, multiple commands, and standardized preprocessing. A short-time Fourier transform or windowed analysis would better represent the nonstationary nature of speech. Perceptual quality measures could also complement RMSE and spectral energy.

6. Conclusion

Under the stated criteria, the soft voice required 193 Fourier terms and the louder, deeper voice required 194 terms to reconstruct the same Alexa command. The experiment therefore found a small variation between the two recordings. The analysis demonstrates how finite Fourier models balance retained spectral information against diminishing gains in reconstruction accuracy. Because the sample was limited to two recordings and the result depended on a fixed 200-term reference, the numerical difference should be treated as an exploratory finding rather than a general rule about speakers.

References

Innovation World. (2025). Parseval's theorem. https://innovation.world/invention/parseval-theorem/

Weisstein, E. W. (2008). Fourier series. MathWorld-Wolfram Web Resource. https://mathworld.wolfram.com/FourierSeries.html