



Original scientific paper

Reimagining Architecture: A Semiotic Study of Sound in Ai-Generated Spatial Design

* Hidayet Softaoğlu

Department of Art, Design, and Architecture, Alanya Alaaddin Keykubat University, Türkiye

E-mail: hidayet.softaoglu@alanya.edu.tr

ORCID: <https://orcid.org/0000-0003-2208-691X>

ARTICLE INFO:

Article History:

Received: 3 April 2025

Revised: 10 July 2025

Accepted: 15 July 2025

Available online: 20 July 2025

Keywords:

Semiotics,
Architectural Theory and Criticism,
Architectural Design,
Artificial Intelligence,
Multi-Sensory Design.

ABSTRACT

This study examines how artificial intelligence (AI) interprets spoken architectural language by analysing vocal features—such as pitch, tone, and magnitude—and translating them into visual representations. Situated within the field of architectural semiotics, the research investigates how sound functions not merely as an acoustic phenomenon but as a symbolic agent in AI-mediated design. Five ambiguous architectural terms (vault, shell, column, plan, story) were recorded in two distinct sentence contexts and vocal styles (neutral vs. expressive). Using the Librosa Python library, pitch range and vocal magnitude were extracted as prosodic features. These metrics informed the construction of emotionally nuanced text prompts for MidJourney, a generative AI model, to produce architectural images reflecting vocal delivery. The results reveal consistent correlations between vocal variation and visual form: high pitch and strong vocal energy led to expressive, fluid, and emotionally charged spaces; lower pitch and stable magnitude generated grounded, monumental, or contemplative structures. These outcomes suggest that vocal expression can serve as a semiotic input in cross-modal AI workflows, where speech acts as both data and design material.

By bridging sound and space, the study expands the semiotic framework of architectural representation and introduces voice as a generative modality in AI-assisted urban and spatial design. The findings support a multi-sensory design paradigm where not only what is said, but how it is said, shapes architectural meaning.

This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 International (CC BY) license.



Publisher's Note:

Journal of *Smart Design Policies* stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

SMART DESIGN POLICIES (2025), 2(1), 107–121.

<https://doi.org/10.38027/smart.v2n1-7>

www.smartdpj.com

Copyright © 2025 by the author(s).

* Corresponding Author

How to cite this article: (APA Style)

Softaoğlu, H. (2025). Reimagining architecture: A semiotic study of sound in AI-generated spatial design. *Smart Design Policies*, 2(1), 107–121. <https://doi.org/10.25034/smart-v2n1-7>

1. Introduction

1.1 Background and Context

Architecture has long served as a medium to communicate meaning, culture, and emotion spatially. In recent years, the intersection of artificial intelligence (AI) and design has opened up new modalities for architectural expression, extending beyond conventional visual and textual domains (Picon, 2010). This paper explores one such emerging modality: the semiotic role of sound, specifically spoken architectural language, in shaping AI-generated visual outputs.

Rooted in the field of semiotics, which studies how signs convey meaning, this research examines how tone, pitch, rhythm, and vocal emphasis influence the interpretation of architectural homographs—words like "vault," "shell," or "plan," which carry multiple meanings depending on their context (Chandler, 2017). While traditional semiotic analysis privileges the visual symbol, this

study extends that analysis into the auditory realm. It asks: How does AI interpret the subtleties of spoken language and translate them into architectural form?

Using a combination of *Librosa* (an audio-processing library) and MidJourney (a text-to-image generation model), this research introduces a novel methodology for converting spoken architectural expressions into images (McFee et al., 2015). By analysing audio features such as pitch range and vocal magnitude and embedding these qualities into prompt-engineered visualisations, the study examines how AI maps acoustic cues to spatial representation. This cross-modal workflow enables a deeper understanding of how AI processes text and interprets the emotive and rhetorical force behind speech (Radford et al., 2021).

The study contributes to architectural semiotics and AI design theory by focusing on ten sentence variations across five architectural terms and linking their vocal qualities to resulting images. It demonstrates that the meaning in architecture, when mediated by AI, is what is said and how it is said. The findings suggest the potential for a multi-sensory design paradigm where voice becomes an architectural tool.

This study addresses the following central research question: How do prosodic features of spoken architectural language—such as pitch, rhythm, and vocal intensity—influence the way AI systems generate architectural imagery? More specifically, it asks whether these vocal attributes function as semiotic signifiers in multimodal AI workflows that interpret language across sound and space.

To answer this, the paper is structured as follows: Section 2 presents a theoretical framework on semiotics, language, and AI. Section 3 introduces the concept of architectural homographs. Section 4 outlines the methodology, combining acoustic signal analysis with AI prompt engineering. Section 5 discusses the visual outcomes and their symbolic interpretations, while Section 6 reflects on the broader implications of sound as a design input. The paper concludes by proposing directions for future multimodal design research.

2. Literature Review / Theoretical Framework

The relationship between language and architecture has long been a focus within the disciplines of semiotics and design theory. Seminal theorists such as Umberto Eco (1976) and Roland Barthes (1977) have emphasised that architecture is not merely functional but operates as a system of signs, a language through which society communicates ideology, identity, and culture. Charles Jencks (1977) further advanced this view by linking post-modern architecture to linguistic plurality, arguing that buildings, like texts, can carry multiple meanings depending on their context and form. This symbolic layering of space and material also aligns with Frampton's (1995) concept of tectonic culture, where construction is not merely technical, but a poetic expression embedded with cultural meaning. It resonates with André Leroi-Gourhan's (1993) view of speech as an extension of gesture, where voice is physical and spatial.

In visual semiotics, Kress and van Leeuwen (2006) developed a grammar of visual design, establishing that images, like spoken or written language, are structured by culturally informed codes. This insight is crucial for understanding AI-generated imagery, as it frames visual outputs not as neutral renderings but as compositions embedded with symbolic meaning.

With the advent of deep learning and generative models, semiotics has expanded into new technological domains. Models such as CLIP (Contrastive Language–Image Pretraining) by OpenAI (Radford et al., 2021) and DALL·E have redefined how machines interpret the relationship between text and image. These developments are underpinned by foundational work in deep learning, which enabled the training of large-scale neural networks to perform generative and interpretive tasks (Goodfellow, Bengio, & Courville, 2016). These capabilities were further expanded through the use of convolutional neural networks (CNNs), enabling deep feature abstraction in visual analysis (Szegedy et al., 2015). These models embed linguistic and visual data in a shared semantic space, allowing them to match or generate images based on textual prompts. While these systems are commonly used for static text input, their potential for incorporating other modalities, such as sound, remains underexplored.

Recent studies in multimodal AI (Padi et al., 2022) have begun to investigate how audio signals, such as prosody, emotion, and intonation, affect machine learning interpretations. In these systems, tone of voice becomes a semiotic variable, conveying emotional or contextual information that complements or overrides textual meaning. These findings suggest that AI models could be trained to understand what is said and how it is said—a fundamental concept in semiotics and communication theory (van Leeuwen, 1999).

Despite this growing interest, there is a notable lack of research on how these capabilities translate to architecture and urban visualisation. Most applications of AI in architecture have focused on generative design, parametric modelling, or material optimisation (Oxman, 2017). As Oxman and Oxman (2014) note, digital design theories, including computational and generative approaches, have significantly reshaped architectural workflows and epistemologies. Very few have addressed how vocal expression might influence architectural form, primarily through symbolic or emotional interpretation.

This study positions itself at the intersection of semiotic theory, multimodal AI, and architectural design. It draws on semiotic frameworks to analyse how sound features, such as pitch and magnitude, can be treated as signifiers (Barthes, 1977; Peirce, 1958). It also applies the affordances of AI image-generation tools—particularly MidJourney—to interpret these auditory signifiers into spatial forms. In doing so, the research contributes to an emerging discourse on cross-sensory translation in design and expands the role of semiotics beyond visual literacy to include acoustic cognition.

2.2 Research Gap and Objectives

The relationship between architecture and artificial intelligence (AI) has been the focus of a growing body of research. However, most studies primarily concentrate on visual or textual modalities (Enjellina et al., 2023). These include areas such as generative design (Coeckelbergh, 2023), parametric modelling (Sage, 2022), originality and creativity in design (Mikalonyté & Kneer, 2022), and utilising textual prompts for image creation (Baran, 2023). In contrast, the role of sound, particularly spoken language and its prosodic elements, has received limited attention in AI-driven design processes.

While multimodal AI systems are starting to utilise audio signals for emotion recognition and user interaction, their potential applications in architectural interpretation and spatial design remain largely unexplored. Additionally, although semiotic theory has long recognised the symbolic significance of language in architecture, the translation of auditory signifiers into spatial forms is still in its early stages.

This study aims to fill this gap by exploring the question: How can AI interpret vocal qualities such as pitch, volume, and rhythm as semiotic indicators of architectural meaning? The goal is to investigate whether and how sound can serve as a generative input in cross-modal workflows that connect speech and spatial design. To achieve this, the study seeks to develop a methodological framework that links voice to visual design through acoustic analysis and AI image generation.

2.3 Contribution and Structure of the Paper

This paper contributes to the growing discourse on multi-sensory design, semiotic AI interpretation, and voice-based spatial cognition by introducing an innovative method for translating verbal architectural expressions into visual representations using Librosa and MidJourney. It extends the field of architectural semiotics beyond visual and textual analysis into the auditory domain, demonstrating how prosodic features of speech influence spatial, symbolic, and emotional outputs in AI-assisted design. The structure of the paper is as follows:

- Section 3 outlines the research methodology, which includes acoustic signal processing, prompt engineering, and visual semiotic analysis.
- Section 4–5 presents the dataset and image results derived from five architectural homographs recorded in dual vocal styles.
- Section 6 discusses the findings with a focus on how voice influences AI spatial interpretations.

- Section 7 concludes with theoretical implications and proposes directions for future research on real-time voice-to-space systems.

3. Materials and Methods

This study employs a qualitative, practice-based research methodology to investigate how the prosodic features of spoken architectural language—such as pitch, vocal magnitude, and rhythm—can impact AI-generated architectural imagery. Given the symbolic and affective nature of the inquiry, a qualitative approach is appropriate for interpreting how meaning is encoded and decoded across sound and spatial representation.

The research consists of four key stages: (1) selection of ambiguous architectural terms, (2) audio recording and feature extraction, (3) prompt engineering and image generation, and (4) semiotic analysis of AI-generated visuals (Figure 2).

3.1 Selection of Ambiguous Architectural Terms

Five polysemous architectural terms were selected for their dual literal and metaphorical meanings: *vault*, *shell*, *column*, *plan*, and *story*. Each term was embedded in two distinct sentence contexts—one neutral and technical, and one expressive and poetic—to modulate semantic connotation. These terms were chosen for their prevalence in architectural discourse and their tonal flexibility in spoken language.

3.2 Audio Recording and Acoustic Feature Extraction

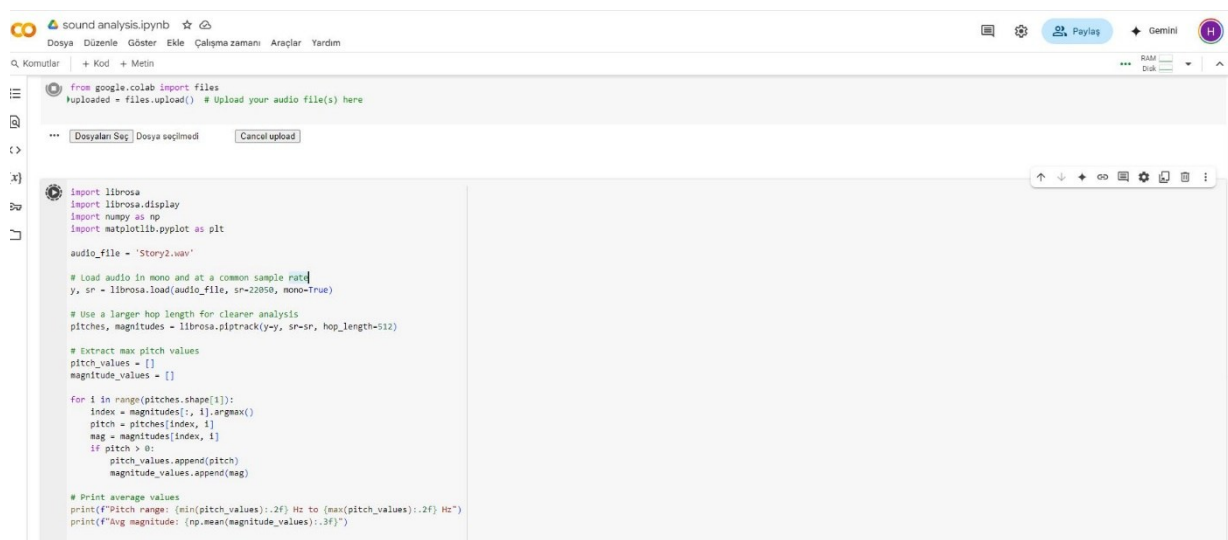
The ten sentences (five terms × two vocal styles) were recorded using the researcher’s voice in a controlled acoustic environment. Emphasis was placed on modulating pitch, stress, and rhythm to express variations in emotional tone. Audio signals were processed using the Librosa Python library in a Google Colab environment (McFee et al., 2015).

Two key acoustic features were extracted:

Pitch range (Hz): Frequency variation linked to emotional intensity.

Average magnitude (dB proxy): Vocal loudness approximated as expressive force.

The analysis was conducted using *librosa.pyin()* (figure 1) for pitch tracking and RMS-based methods for magnitude measurement. A hop length of 512 was used to ensure clarity and resolution in feature extraction.



```
from google.colab import files
uploaded = files.upload() # Upload your audio file(s) here

... [Dosyaları Seç] Dosya seçilmedi [Cancel upload]

import librosa
import librosa.display
import numpy as np
import matplotlib.pyplot as plt

audio_file = 'Story2.wav'

# Load audio in mono and at a common sample rate
y, sr = librosa.load(audio_file, sr=22050, mono=True)

# Use a larger hop length for clearer analysis
pitches, magnitudes = librosa.piptrack(y=y, sr=sr, hop_length=512)

# Extract max pitch values
pitch_values = []
magnitude_values = []

for i in range(pitches.shape[1]):
    index = magnitudes[:, i].argmax()
    pitch = pitches[index, i]
    mag = magnitudes[index, i]
    if pitch > 0:
        pitch_values.append(pitch)
        magnitude_values.append(mag)

# Print average values
print(f"Pitch range: {min(pitch_values):.2f} Hz to {max(pitch_values):.2f} Hz")
print(f"Avg magnitude: {np.mean(magnitude_values):.3f}")
```

Figure 1. The acoustic analysis was conducted using a Python script in Google Colab based on the *Librosa* library (McFee et al., 2015). The code extracted pitch contours and average magnitude to represent emotional intensity and vocal emphasis in speech recordings (by the Author).

3.3 Prompt Engineering and Image Generation (MidJourney)

The extracted acoustic features were used to craft prompts for MidJourney, an AI-based text-to-image generation model. Rather than literal transcriptions, prompts were emotionally expressive, shaped by vocal performance data. High-pitched, intense vocal inputs produced prompts describing dynamic, radiant architecture, while low-pitched, calmer voices inspired grounded and serene designs.

Visualisations were generated using MidJourney v6 with the /imagine command and the --ar 16:9 aspect ratio for stylistic consistency. Ten images were created in total.

All images used in this study were generated exclusively by the author using original voice recordings and prompt design via MidJourney. No pre-trained third-party visual datasets or externally sourced prompts were used. This ensures full creative authorship and supports the originality of the visual material presented.

3.4 Visual Semiotic Analysis

AI-generated images were analysed using semiotic frameworks, particularly Kress and van Leeuwen's grammar of visual design (2006) and Barthes' (1977) connotation theory. Each image was examined for:

- Spatial composition and material articulation
- Emotional or symbolic resonance
- Mood, rhythm, and light-texture interplay

These interpretations were then compared with the original acoustic profiles to identify correlations between voice and form, revealing how AI responds to vocal nuance as a design parameter.

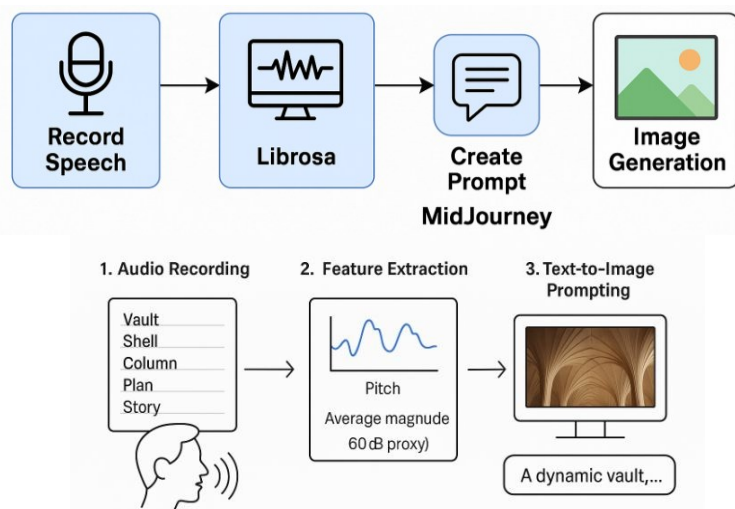


Figure 2. The diagram shows the methodological map. The first two different contexts were chosen for each architectural term: *vault*, *shell*, *column*, *plan*, and *story*. The author recorded each sentence in their own voice. The second step was analysing them in *Librosa*. The third step was using their Hz and dB proxy value to tailor prompts in MidJourney. The fourth step was generating images and analysing them (by the Author).

4. Architectural Homographs

Architectural language is replete with homographs—words with multiple meanings based on context, tone, and discipline. This semantic richness makes architectural vocabulary particularly fertile ground for a study in semiotic variation. Many architectural terms are polysemous, encompassing literal structural definitions and metaphorical or interdisciplinary usages (Chandler, 2017; Eco, 1984).

Table 1 below illustrates a broader set of architectural homographs and their possible meanings:

Table 1. Architectural Terms with Contextual or Tonal Ambiguity (Developed by Author).

Term	Possible Meanings
Vault	Curved ceiling (architecture), secure room (bank), leap/move (verb)
Shell	Exterior structure, exoskeleton (biology-inspired), explosion casing
Column	Structural support, typographic layout, procession or queue
Bay	Spatial division (between columns or windows), body of water, to bark (verb)
Plan	Architectural drawing, intention/strategy (plan of action)
Frame	Structural skeleton, visual boundary (camera shot), manipulation ("frame someone")
Pier	Support column (in bridges), waterfront structure, part of a face in Gothic windows
Buttress	Support structure (architecture), metaphor for support (argument, idea)
Elevation	Architectural drawing, height above sea level, spiritual/metaphorical rising
Section	Architectural cut-through, portion/division (e.g., of a city or text)
Axis	Central line in plan, ideological/political grouping ("axis of evil")
Deck	Outdoor platform, surface level (ships), to decorate (verb)
Roof	Cover of a building, metaphor for limit ("hit the roof")
Footing	Structural base, beginning of a walk or journey, metaphorical grounding
Story	Floor level of a building, narrative or fictional account
Span	Distance between supports, time duration, extent of influence
Truss	Structural framework, to tie or bind (verb), medical support
Mass	Large volume (form), religious gathering, physical weight or density
Program	Building function or use case, sequence of actions or code, broadcast content

Each of these words, when spoken, can be modulated in tone or stress to reflect a different *semiotic intention*. For instance:

- Saying "bay" in a calm, downward intonation may suggest a coastal context; staccato and sharpness may suggest a barking dog, confusing AI if used literally.
- "Plan" in a soft, speculative tone could suggest a conceptual strategy, while a flat, technical pronunciation might evoke a blueprint.

These subtleties are lost in traditional AI systems that use text prompts only. Still, they can unlock far richer architectural representations when combined with audio data—and interpreted through a semiotic lens (van Leeuwen, 1999).

For this study, five homographs were selected based on their widespread use, semantic flexibility, and relevance in both literal and conceptual architectural discourse. The chosen terms—*vault*, *shell*, *column*, *plan*, and *story*—represent a cross-section of architectural vocabulary that shifts meaning depending on how they are articulated and contextualised. These terms were recorded in two distinct spoken sentence contexts to test the hypothesis that pitch and vocal emphasis changes alter AI-generated architectural interpretations (McFee et al., 2015; Radford et al., 2021).t

5. From Voice to Visual: Translating Acoustic Features into Architectural Images

This section details how spoken architectural terms were converted into visual representations using artificial intelligence tools. After the audio recording phase, each sentence was processed in Google Colab using Python's *Librosa* library to extract measurable acoustic features—specifically, pitch range (Hz) and average vocal magnitude. These features were interpreted semiotically to reflect emotional tone and rhetorical variation (McFee et al., 2015).

Five semantically rich architectural terms were selected to examine how vocal tone influences architectural meaning: *vault*, *shell*, *column*, *plan*, and *story*. Each term carries multiple definitions depending on context—e.g., "vault" can refer to a structural arch or a secure space like a bank vault.

Two distinct sentences were recorded for each term to highlight these contextual differences. In each case, vocal modulation—tone, rhythm, pitch, and inflexion was adjusted to suggest either a literal or metaphorical interpretation (Gaver, 1993).

The recorded sentences included:

Vault: "The vault of the cathedral was intricately designed, creating a sense of openness and grandeur."

Shell: "The building's shell was completed, but the interior is still under construction."

Column: "The columns of the ancient temple were massive, standing as a testament to the civilisation."

Plan: "The architect presented the plan for the new civic centre, featuring modern lines and sustainable materials."

Story: "The building has five stories, each with floor-to-ceiling windows for natural light."

Each sentence was delivered in two distinct tonal styles:

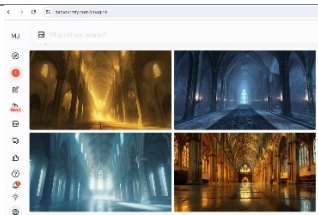
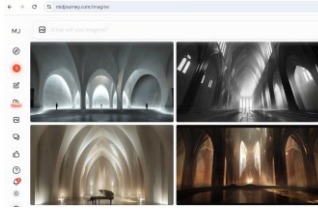
- A neutral, technical tone for literal architectural meaning
- A poetic or emotionally inflected tone suggests symbolic or metaphorical meaning.

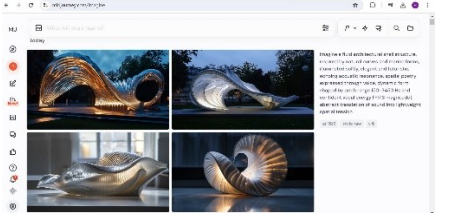
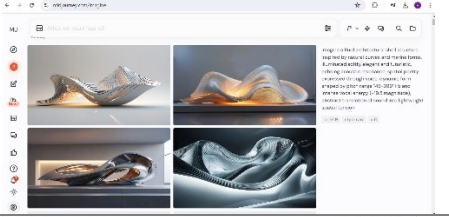
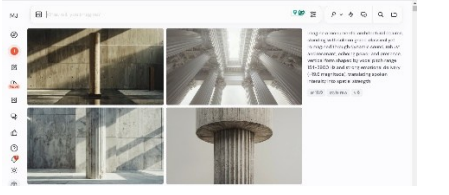
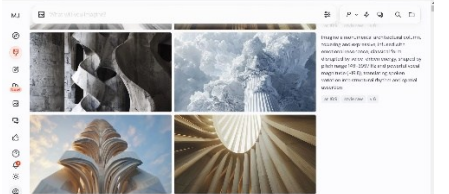
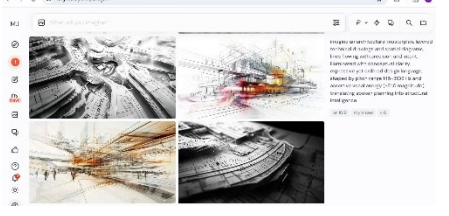
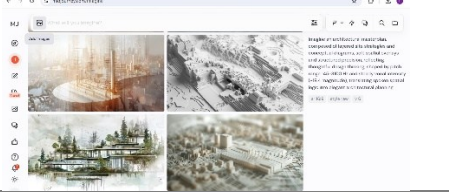
These recordings were then analysed using *Librosa* to extract pitch and magnitude data, which were treated as signifiers within a semiotic framework. This data informed the construction of semantically enriched prompts for MidJourney, an AI-powered text-to-image generator (Kress & van Leeuwen, 2006).

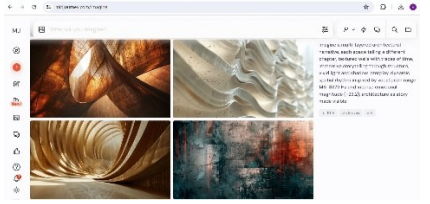
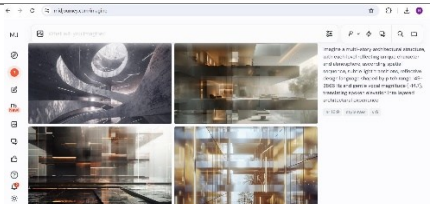
Rather than inserting raw audio files, the extracted acoustic metrics were translated into textual descriptions. Higher pitch and vocal magnitude led to prompts emphasising verticality, brightness, formal complexity, and emotional tension. In contrast, lower values produced prompts describing grounded, enclosed, or contemplative spaces. All prompts were submitted using MidJourney's/imagine command with --v 6 and --ar 16:9 flags to maintain output consistency (Radford et al., 2021).

In total, ten visual outputs were generated—two per term. Each was documented alongside its corresponding sentence, acoustic metrics, and MidJourney prompt. A visual-semiotic analysis followed, assessing how vocal delivery shaped spatial form, texture, atmosphere, and symbolic resonance (Barthes, 1977; van Leeuwen, 1999). The full dataset and visual comparisons are provided in Table 2.

Table 2: Shows vocalising the same sentence can give different semiotic results in terms of pitch and magnitude (all images generated by the author in Midjourney v6).

Word	Sound	Pitch Range	Avg Magnitude	Interpretation	Prompt (MidJourney)	Image
Vault	1	149.07 - 3898.48 Hz	17.894	Wide pitch range with strong vocal energy suggests dramatic, expressive delivery with reverence.	a vaulted gothic cathedral, softly illuminated, grand yet serene atmosphere, echoing silence, high arching ceiling, sacred spatial feeling, Pitch: 182–216 Hz, Magnitude: 0.035 --v 6 --ar 16:9	
Vault	2	153.67– 3153.68 Hz	18.280	Similar to Sentence 1 — wide pitch with strong vocal intensity; implies spatial grandeur and emotional resonance.	a vaulted gothic cathedral, softly illuminated, grand yet serene atmosphere, echoing silence, high arching ceiling, sacred spatial feeling, infused with sonic intensity, dramatic vocal energy interpreted through architecture, dynamic spatial tension reflecting pitch variation from 153.67 Hz to 3153.68 Hz, assertive emotional resonance inspired by high vocal magnitude (~18.280), a	

					fusion of sound and space -- v 6 --ar 16:9	
Shell	1	150.74– 3450.79 Hz	17.386	Broad tonal range with confident delivery; suggests flowing, poetic, nature-inspired design.	a fluid architectural shell structure, inspired by natural curves and marine forms, illuminated softly, elegant and futuristic, echoing acoustic resonance, spatial poetry expressed through voice, dynamic form shaped by pitch range 150–3450 Hz and confident vocal energy (~17.3 magnitude), abstract translation of sound into lightweight spatial tension -- v 6 --ar 16:9	
Shell	2	146.95– 3991.13 Hz	19.495	Very expressive tone with intense vocal power; suggests a bolder, more charged shell form.	(Same prompt, adapted with pitch/magnitude values)	
Column	1	154.91– 3900.60 Hz	19.613	Strong, expressive vocal quality; evokes classical power and monumental architectural rhythm.	a monumental architectural column, standing with solemn grace, classical yet reimagined through dynamic sound, robust and resonant, echoing power and presence, vertical form shaped by vocal pitch range 154–3900 Hz and strong emotional delivery (~19.6 magnitude), translating spoken intensity into spatial strength --v 6 -- ar 16:9	
Column	2	149.56– 3997.94 Hz	19.804	Highly dynamic delivery, slightly more charged than Sentence 1; suggests tension and vertical force.	(Same prompt, with new pitch/magnitude values)	
Plan	1	145.75– 3108.30 Hz	20.986	Extremely assertive and energetic; suggests confident, technical spatial reasoning.	an architectural masterplan, layered technical drawings and spatial diagrams, lines flowing with precision and intent, illuminated with conceptual clarity, expressive yet ordered design language, shaped by pitch range 145–3108 Hz and assertive vocal energy (~21.0 magnitude), translating spoken planning into structural intelligence -- v 6 --ar 16:9	
Plan	2	145.88– 3100.32 Hz	19.429	Slightly softer than Sentence 1, reflective and calculated delivery; evokes conceptual clarity.	(Same prompt, adapted with lower magnitude)	


Story	1	146.59–3379.66 Hz	23.213	Very expressive and loud; suggests storytelling architecture filled with emotional movement.	a multi-layered architectural narrative, each space telling a different chapter, textured walls with traces of time, immersive storytelling through structure, vivid light and shadow interplay, dynamic spatial rhythm inspired by vocal pitch range 146–3379 Hz and intense emotional magnitude (~23.2), architecture as story made visible --v 6 --ar 16:9	
Story	2	149.57–3866.24 Hz	14.759	Wide pitch but calm delivery; evokes thoughtful spatial layering and gentle transitions.	a multi-story architectural structure, with each level reflecting unique character and atmosphere, ascending spatial sequence, subtle light transitions, reflective design language shaped by pitch range 149–3866 Hz and gentle vocal magnitude (~14.7), translating spoken elevation into layered architectural experience --v 6 --ar 16:9	




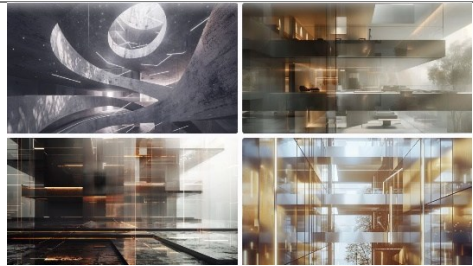
The following comparative tables (Table3) covering *vaults*, *shells*, *columns*, *plans*, and *stories* illustrate how variations in vocal attributes, specifically pitch range and vocal magnitude, influence architectural outcomes generated by MidJourney AI. Each table presents a side-by-side comparison of two images created using identical prompts, differing only in acoustic input features.

Organised across five interpretive dimensions; spatial form, scale and hierarchy, light and material, symbolic resonance, and atmospheric mood, these tables reveal how the AI system consistently maps lower pitch and stable magnitude to monumental, grounded, and structured environments. In contrast, higher pitch and dynamic vocal energy often produce fluid, expressive, and emotionally charged forms.

The expressive impact of acoustic modulation becomes evident in each case, revealing how variations in pitch and vocal intensity function as affective and symbolic triggers within AI-generated architectural imagery. In the *vault* examples, the transition from solemn, grounded arches to theatrical and distorted spatial compositions illustrates how vocal tone modulates spatial tension and emotional charge. The *plan* table demonstrates how diagrammatic logic and urban rhythm are directly shaped by the speaker's sense of urgency or calm, resulting in divergent urban patterns that reflect either structured pacing or chaotic expansion. The *shell* outputs exemplify how shifts in vocal magnitude correlate with spatial atmosphere: calm delivery yields meditative, enclosed geometries, while more dynamic expression produces fluid, biomorphic enclosures. In the *column* and *story* examples, expressive vocal range appears to shift the very architectural typology, evoking, respectively, either civic monumentality or mythologically charged narrative environments. Collectively, these comparative analyses reveal that acoustic prosody, far from being a superficial input, operates as a latent semiotic code that informs form, mood, and symbolic resonance. Through voice, architecture becomes not just constructed, but emotionally narrated.

Table 3: Shows how pitch range and vocal magnitude change & affect the design result (all images generated by the author in Midjourney v6).

Aspect (Vault)	 <p>(149.07–3898.48 Hz / ~17.894 Mag.)</p>		 <p>(153.67–3153.68 Hz / ~18.280 Mag.)</p>	
Spatial Form	Symmetrical, elongated arches; smooth, continuous vaults		Organic, expressive vaulting; fluid transitions between surfaces	
Scale & Hierarchy	Classical proportions; consistent vertical rhythm		Hybrid scale; spatial layering implies harmonic complexity	
Light & Material	Gentle diffusion, soft palette (blues and golds); matte textures		Translucent glows, polished textures; chiaroscuro interplay of tone & light	
Symbolic Resonance	Stillness, serenity, spiritual echo		Sonic embodiment of divine power; fusion of liturgical and performative space	
Architectural Mood	A contemplative sanctuary		A cathedral-concert hybrid — sacred yet performative, immersive and alive	
Aspect (Shell)	 <p>150–3450 Hz / ~17.3 Mag.</p>		 <p>146–3991 Hz / ~19.5 Mag.</p>	
Spatial Form	Lightweight arcs, softly segmented curves; balanced organic structure		Fluid, biomorphic complexity with flowing intensity	
Scale & Hierarchy	Horizontal orientation, gentle human-scale presence		Dynamic layering, expanded, immersive sculptural volume	
Light & Material	Translucent shell-like surfaces, soft interior glows		Polished metallic skins, glowing interiors, and high contrast light play	
Symbolic Resonance	Acoustic calm, marine elegance, visual-poetic stillness		Expressive energy, sonic turbulence, dynamic lyricism	
Architectural Mood	Quiet meditation pavilion; a spatial poem		Futuristic performance sculpture; vibrant and bold	
Aspect (Column)	 <p>(154–3900 Hz / ~19.6 Mag.)</p>		 <p>(149–3997 Hz / ~19.8 Mag.)</p>	
Spatial Form	Angular, fluted columns; rigid geometry		Curved, baroque, morphing forms	
Scale & Hierarchy	Structured, classical hierarchy		Amplified scale, disrupted hierarchy	
Light & Material	Sharp light, concrete/stone textures; austere illumination		Diffused light, surreal/polished textures	

Symbolic Resonance	Power, solemnity, permanence — like a deep, solemn voice	Emotional, expressive, mythic — like an impassioned or rising voice
Architectural Mood	Civic monument or sacred memorial	Mythic cathedral or emotionally resonant dream space
Aspect (Plan)	 <p>(145–3108 Hz / ~21.0 Mag.)</p>	 <p>(145–3100 Hz / ~19.4 Mag.)</p>
Diagrammatic Form	Layered urban networks, axial bursts, high-density vector logic	Soft overlays, ecological zoning, recursive territorial stratification
Planning Scale & Intensity	Urban megastructures; rapid data-like energy	Regional/environmental scope; long-range balance
Visual Language	Sharp technical lines, saturated motion highlights, grid compression	Washed tones, organic textures, soft hierarchy
Symbolic Resonance	Clarity, strategy, verbal command as architectural imprint	Thoughtfulness, adaptability, wisdom through calm articulation
Cognitive Mood	Strategic urgency; controlled chaos	Meditative foresight; slow and steady structure of thought
Aspect (Story)	 <p>(146–3379 Hz / ~23.2 Mag.)</p>	 <p>(149–3866 Hz / ~14.7 Mag.)</p>
Narrative Form	Expressive curves, layered wood and fabric forms	Linear stacking, transparent layering
Spatial Language	Sensual, flowing, immersive gestures	Ascending sequence, restrained articulation
Light & Material	Rich grains, warm shadows, heavy contrast	Glassy gradients, subtle glows, minimal contrasts
Symbolic Resonance	Emotional storytelling, memory and presence	Quiet elegance, introspective progression
Atmospheric Mood	Theatrical and emotive spatial drama	Calm, intellectual, and contemplative ascent

6. Findings and Discussion: Sound as a Semiotic Agent in AI-Driven Architectural Design

This study examined how vocal characteristics, specifically pitch range, vocal magnitude, rhythm, and tone, are interpreted by AI (specifically MidJourney v6) when generating architectural and urban design outputs. The results (Table 4) demonstrate that AI does not simply convert inputs into visuals but instead conducts a multimodal synthesis in which sound functions as a symbolic language, influencing material articulation, spatial composition, and atmospheric mood (van Leeuwen, 1999; Barthes, 1977).

Table 4: Acoustic Observations and Theoretical Implications (by Author).

Design Terms	Observed Pattern	Theoretical Implication
1 Vault: Spatial Tension and Emotional Charge	Low pitch produces grounded arches; high pitch results in theatrical distortion.	AI interprets vocal tension as spatial drama, suggesting voice can encode emotional tension in form.
2 Shell: Biomorphic Calm vs Dynamic Enclosure	Calm delivery yields meditative curves; dynamic tone creates flowing, aggressive shells.	Prosodic force maps onto spatial enclosure logic—suggesting AI visualises intensity as enclosure morphology.
3 Column: Monumentality and Expressive Verticality	Stable tone results in civic monumentality; expressive vocal range generates mythic forms.	Expressive range in voice drives typological shifts, indicating affect as an architectural classifier.
4 Plan: Urban Logic and Rhythmic Modulation	Urgent tone leads to sprawling, diagrammatic patterns; calm tone generates coherent layouts.	AI uses rhythm and vocal urgency to spatialise pacing and density, showing speech tempo as urban logic.
5 Story: Narrative Emotion and Atmospheric Layering	High-pitched inputs result in dramatic layering and emotional visuals; low-pitched voices yield serene, vertical structures.	Emotional tone in voice constructs narrative atmosphere, revealing voice as a symbolic-organising principle.

Spatial Form and Pitch-Magnitude Translation

Across the dataset, pitch and magnitude variations showed consistent effects on spatial geometry: Lower to moderate pitch with steady magnitude (e.g., columns at 154–3900 Hz, ~19.6 magnitude) produced monumental, orthogonal forms conveying structural integrity and calm authority. Higher pitch with dynamic magnitude (e.g., shells at 146–3991 Hz, stories at ~23.2) yielded fluid, expressive geometries, evoking motion, disruption, and emotional intensity. These findings suggest that pitch is interpreted as a cue for curvature and formal complexity, while magnitude signals scale and spatial pressure. These interpretations align with auditory semiotics, where high pitch often connotes urgency or energy, and low pitch suggests solemnity or stability (Gaver, 1993).

Symbolic Resonance and Acoustic Semiotics

The analysis revealed that tone and rhythm operate as symbolic signifiers in AI-generated designs: Assertive, forceful delivery produced outputs with sharper geometries, denser spatial hierarchies, and compressed compositions, evoking a sense of urgency or control (e.g., urban plans). A soft or contemplative tone resulted in translucent layering, open forms, and diffused material transitions, suggesting introspection, spiritual presence, or memory (e.g., vaults, stories). This demonstrates that AI translates not just semantic content but vocal expression into architectural symbolism, encoding affective and cultural registers into spatial form.

Atmosphere as a Function of Voice

Atmospheric qualities such as light diffusion, surface texture, and spatial rhythm closely followed vocal variation:

High emotional intensity was associated with dramatic lighting, textured surfaces, and immersive layering.

Low vocal modulation correlated with soft illumination, minimal articulation, and calm, contemplative spatial mood. These outcomes suggest that vocal attributes serve not only to define form but also to guide environmental atmosphere, making voice a generative agent in multi-sensory

design. This reflects Norberg-Schulz's (1980) idea that architectural space is inseparable from atmosphere—a "genius loci" shaped by material, light, and mood.

Urban and Diagrammatic Semiotics

In more abstract configurations such as site plans the AI also responded sensitively to acoustic signals: High-magnitude input generated dense, technical master plans suggestive of urgency and urban complexity.

Lower-magnitude delivery led to broader, ecologically informed layouts prioritising spatial balance and rhythm. This underscores how vocal tone can convey planning logic, mirroring how human speech reflects intention and scale in architectural discourse (Picon, 2010) and supporting Schön's (1992) notion of design as reflective conversation, where designers iteratively respond to situational cues—in this case, vocal expression.

AI and the Mediation of Sound as Semiotic Code

Viewed through a semiotic lens, AI emerges as a mediator of phonetic signifiers, translating vocal tone, rhythm, and emphasis into architectural meaning. The system does not merely visualise linguistic input but responds to how speech is performed:

Stress and rhythm influence spatial pacing and formal repetition.

Pitch and tone encode affective, ideological, and symbolic cues. By mediating these variables, AI establishes a cross-modal design logic where voice becomes a primary design medium. This rhythmic alignment between speech and space mirrors principles of emergent order found in complex systems (Strogatz, 2003).

These findings form the analytical backbone for the conclusion, where implications for multisensory architectural design and AI-driven urban planning are synthesised.

7. Conclusions

This study examined how vocal attributes, specifically pitch, magnitude, tone, and rhythm, function as semiotic agents in AI-driven architectural design. By combining acoustic analysis through *Librosa* with image generation via MidJourney, it demonstrated that vocal variation influences not only the visual qualities of architectural outputs but also their spatial atmosphere, material expression, and symbolic resonance.

A key contribution of this research lies in its extension of architectural semiotics beyond the visual realm, into an auditory dimension. While most AI-driven design systems rely on denotative, text-based prompts, this study has shown that prosodic elements of speech, such as emphasis, cadence, and inflection, can significantly affect the outcomes of architectural visualisation. Specifically, the findings indicate that:

- High-pitched and dynamic vocal deliveries are associated with expressive, emotionally charged architectural forms and intense atmospheric lighting.
- Low-pitched and steady vocal inputs tend to generate monumental, structured, or contemplative environments.

These consistent relationships between voice and visual-spatial expression suggest that sound can act as a generative tool in architectural design. Rather than functioning as mere input, voice emerges as a medium of emotional, formal, and spatial coding that AI systems can interpret and transform into meaningful architectural representations.

Undoubtedly, AI was discussed for its great potential, especially in the early stages of architectural design (Vissers-Similon, 2024). This research reframes the role of AI in the design process. Instead of viewing AI as a neutral rendering tool, the findings position it as an active interpreter of symbolic and phonetic cues. These reframing challenges traditional separations between language and space, or emotion and form, and support a more integrated understanding of how design might be influenced by vocal expression.

Although the scope of this study was intentionally limited to a specific set of terms and controlled audio variations, it opens several potential directions for future research. For instance, real-time systems could be developed that translate spoken architectural intent into dynamic design feedback.

Cross-cultural analyses may reveal how different prosodic patterns influence AI interpretation of spatial form. Additionally, incorporating multimodal AI systems that combine audio, text, and image could enhance the contextual richness and emotional nuance of generative outputs.

In conclusion, the study affirms that sound is not merely an atmospheric accessory but a meaningful semiotic input in architectural communication. By exploring how AI perceives and reinterprets voice, designers may begin to engage more effective, embodied, and multisensory approaches to space-making.

Ethical and Creative Declaration

All images presented in this article were generated exclusively by the author using original voice recordings and manually crafted prompts with MidJourney v6. No third-party datasets, external collaborators, or pre-trained visuals were used. The acoustic data was self-recorded, and all AI-generated content was ethically produced within a controlled, author-led design process. This ensures full creative authorship and avoids any copyright or intellectual property ambiguities.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Conflicts of Interest

The Author(s) declare(s) that there is no conflict of interest.

Data availability statement

The datasets generated and analysed during the current study are not publicly available but are available from the author on reasonable request.

Institutional Review Board Statement

Not applicable.

CRedit author statement:

Hidayet Softaoğlu: Conceptualisation, Methodology, Investigation, Visualisation, Software, Formal analysis, Writing – Original Draft, Writing – Review & Editing, Data Curation, Resources, Project Administration. The author has read and agreed to the published version of the manuscript.

References

- Baran, M. (2023). *Artificial, Intelligent, Architecture*. ORO Editions.
- Barthes, R. (1977). *Image, music, text* (S. Heath, Trans.). Fontana Press.
- Enjellina & Beyan, E. V. P., & Rossy, A. G. C. (2023). A review of AI image generator: Influences, challenges and future prospects for architectural field. *Journal of Artificial Intelligence in Architecture*, 2(1), 53–65. <https://doi.org/10.24002/jarina.v2i1.6662>
- Chandler, D. (2017). *Semiotics: The basics* (3rd ed.). Routledge.
- Coeckelbergh, M. (2023). The work of art in the age of AI image generation: Aesthetics and human technology relations as process and performance. *Journal of Human-Technology Relations*, 1(1). <https://doi.org/10.59490/jhtr.2023.1.7025>

- Eco, U. (1976). *A theory of semiotics*. Indiana University Press.
- Eco, U. (1984). *Semiotics and the philosophy of language*. Indiana University Press.
- Frampton, K. (1995). *Studies in tectonic culture: The poetics of construction in nineteenth and twentieth century architecture*. MIT Press.
- Gaver, W. W. (1993). What in the world do we hear? An ecological approach to auditory event perception. *Ecological Psychology*, 5(1), 1–29. https://doi.org/10.1207/s15326969eco0501_1
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
- Jencks, C. (1977). *The language of post-modern architecture*. Rizzoli.
- Kress, G., & van Leeuwen, T. (2006). *Reading images: The grammar of visual design* (2nd ed.). Routledge. <https://doi.org/10.4324/9780203619728>
- Leroi-Gourhan, A. (1993). *Gesture and speech* (A. Bostock Berger, Trans.). MIT Press.
- McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in Python. In *Proceedings of the 14th Python in Science Conference* (pp. 18–25). <https://doi.org/10.25080/Majora-7b98e3ed-003>
- MidJourney. (2022). *MidJourney AI: A generative model for creating images from text prompts*.
- Mikalonytė, E. S., & Kneer, M. (2022). Can artificial intelligence make art? Folk intuitions as to whether AI-driven robots can be viewed as artists and produce art. *ACM Transactions on Human-Robot Interaction*, 11(4), 1–19. <https://doi.org/10.1145/3530875>
- Norberg-Schulz, C. (1980). *Genius loci: Towards a phenomenology of architecture*. Rizzoli.
- Oxman, R., & Oxman, R. (Eds.). (2014). *Theories of the digital in architecture*. Wiley.
- Oxman, R. (2017). Thinking difference: Theories and models of parametric design thinking. *Design Studies*, 52, 4–39. doi: 10.1016/j.destud.2017.06.001
- Padi, S., Sadjadi, S. O., Manocha, D. & Sririam, R. D. (2022). Multimodal Emotion Recognition Using Transfer Learning from Speaker Recognition and BERT-Based Models. *The Speaker and Language Recognition Workshop (Odyssey 2022)*, 407-414. <https://doi.org/10.21437/Odyssey.2022-57>
- Peirce, C. S. (1958). *Collected papers of Charles Sanders Peirce* (Vols. 1–8, C. Hartshorne, P. Weiss, & A. W. Burks, Eds.). Harvard University Press.
- Picon, A. (2010). *Digital culture in architecture: An introduction for the design professions*. Birkhäuser.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *arXiv*. <https://doi.org/10.48550/arXiv.2103.00020>
- Sage, M. F. (2022). *Architecture in high resolution*. ORO Editions.
- Schön, D. A. (1992). *The reflective practitioner: How professionals think in action*. Routledge. <https://doi.org/10.4324/9781315237473>
- Strogatz, S. H. (2003). *Sync: How order emerges from chaos in the universe, nature, and daily life*. Hachette Books.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–9). <https://doi.org/10.1109/CVPR.2015.7298594>
- van Leeuwen, T. (1999). *Speech, music, sound*. Macmillan. <https://doi.org/10.1007/978-1-349-27700-1>
- Vissers-Similon, E., Dounas, T., & De Walsche, J. (2024). Classification of artificial intelligence techniques for early architectural design stages. *International Journal of Architectural Computing*, 23(2), 387–404. <https://doi.org/10.1177/14780771241260857>