The “What” Space
Prosodic Variability and Affective Virtual Environments

Jorge Forero
ITI/LARSyS - INESC TEC
Faculdade de Engenharia
Universidade do Porto
jfforero@ludique.cl

Mónica Mendes
ITI/LARSyS
Faculdade de Belas-Artes
Universidade de Lisboa
monicamendes@edu.ulisboa.pt

Gilberto Bernardes
INESC TEC
Faculdade de Engenharia
Universidade do Porto
gba@fe.up.pt

Abstract

What? What! What… How Many Ways to Say the Word What.

The “What” Space is a research project that investigates the emotional variability and expressive potential of spoken language. Focusing on the word “what,” the project examines how shifts in emotional prosody shape its interpretation. Drawing from the IEMOCAP database, we identified 54 instances of “what,” each annotated with categorical emotion labels and continuous values in the valence–arousal–dominance (VAD) space.

To explore these variations, we developed a series of multimodal visualizations: an interactive 3D visualization in Unity3D, network-based structures in Blender, a real-time speech-to-texture web application, and an immersive virtual reality installation hosted on Onland.io.

These experiences allow users to navigate a spatialized soundscape of “what” utterances and directly experience their emotional diversity.

This work contributes to affective computing, speech emotion recognition, and human–computer interaction by proposing novel frameworks for interpreting and visualizing emotional expression through speech. By revealing the rich affective range embedded in a single word, The “What” Space underscores the role of prosody in communication and demonstrates the potential of speech-based affective virtual environments.

Introduction

Affective virtual environments represent an emerging intersection between emotion recognition systems and dynamically adaptive virtual spaces (Forero et al., 2023). These systems are typically composed of two interdependent components: an emotion recognition module and a virtual environment generator (Pinilla et al., 2021a).

Emotion recognition techniques encompass multiple modalities, including facial expression analysis (Ekman & Friesen, 1978), body movement interpretation (Camurri et al., 2003), and biosignal analysis such as galvanic skin response, heart rate variability, respiratory patterns, and electroencephalographic signals (Picard, 1997; Cowie et al., 2001). Among these modalities, speech-based emotion recognition stands out as particularly suitable for naturalistic interactions, especially when Virtual Reality headsets are utilized and electrophysiological measurements are impractical or unavailable.

Speech inherently embodies dual modalities through its symbolic (semantic content) and acoustic (prosodic features) characteristics. Consequently, spoken language can be analyzed through textual sentiment analysis or acoustic feature extraction, each approach offering complementary insights into emotional expression (Scherer, 2003; El Ayadi et al., 2011).

A particularly intriguing phenomenon arises when semantically neutral utterances, such as the word "What", are evaluated solely through their prosodic expression. These ostensibly neutral words, when delivered with emotionally charged intonations, reveal critical points of affective variability, interpretive complexity, and aesthetic significance.

The “What” Space project addresses this phenomenon by investigating the wide emotional spectrum that a single ambiguous vocal utterance—"What?"—can convey. By focusing on the dynamic prosodic qualities inherent in such minimal linguistic expressions, the project offers an innovative framework for exploring emotional variability and contributes to a deeper understanding of affective dimensions in human-computer interactions.

Literature Review

Affective virtual environments can be traced back to the broader field of intelligent virtual environments, first reported by Aylett and Cavazza (2001). Since then, they have evolved significantly through advances in automatic emotion recognition and generative computer graphics.

For instance, Misha Sra et al.'s Auris project demonstrates how music-based affective cues can shape immersive virtual worlds, translating the emotional nuances of songs into visual elements such as geometry, texture, and color (Sra et al., 2017).

Building on this idea, Forero et al. (2022) present Emotional Machines, an interactive installation that generates affective virtual environments in real time by analyzing users' acoustic and semantic expressions. Using multimodal speech-based emotion recognition, the system continuously adapts its visual landscape in response to the user's emotional state, forming a dynamic feedback loop between emotional input and audiovisual output.

Expanding on the synthesis of emotion and image, Patil et al. (2023) employ Generative Adversarial Networks (GANs) and the VQGAN+CLIP framework—which leverages visual and textual representations to guide image generation—to convert spoken emotional expressions into personalized visual artworks.

Similarly, Misra et al. (2024) introduce the FRIDA robotic system, which interprets emotional tones in audio to guide a robotic painter, translating affective vocal input into expressive graphical forms.

Methodology

We used the IEMOCAP (Interactive Emotional Dyadic Motion Capture) database, which contains multimodal recordings with emotional annotations (Busso et al., 2008). The database comprises approximately 12 hours of recorded interactions between professional actors, organized into five sessions with male–female dyads. The recordings include audio, video, textual transcripts, and emotion labels, making it a valuable resource for speech emotion recognition (SER).

Within the IEMOCAP dataset, we searched for utterances marked as emotionally ambiguous—specifically, instances where identical semantic content was delivered with different emotional expressions or interpretations. Through this analysis, we identified the most consistently ambiguous utterance: the single word "what," which appeared across contexts associated with all major emotional categories present in the corpus, including sadness, happiness, anger, fear, disgust, surprise, frustration, and neutrality.

We retained both the categorical emotion labels and the corresponding dimensional emotion scores (Valence–Arousal–Dominance, VAD) provided in the dataset. In total, we filtered 54 distinct occurrences of the word "what," each labeled with one of the corpus's emotion categories and mapped within the three-dimensional VAD space. The most frequent emotion category was anger (11 instances), followed by frustration (10), surprise (9), and neutrality (6). Other categories included excitement (7), sadness (4), happiness (1), and fear (1).

Results

We developed four creative strategies for data-driven audiovisualization to reveal the emotional ambiguity embedded in the utterance "what." While each strategy yields an independent result, the creative process unfolded sequentially, with each proposal building upon the outcomes of the previous one.

Say What? — Interactive Visualization in VAD Space[1]

We created an interactive 3D visualization in Unity3D to explore the emotional landscape of the word "what" within the three-dimensional Valence–Arousal–Dominance (VAD) space. Each utterance is represented as a sphere, positioned according to its VAD coordinates and color-coded based on its corresponding categorical emotion (see Image 1). Users can freely orbit the camera around a custom reference frame, enabling intuitive spatial exploration of emotional variation.

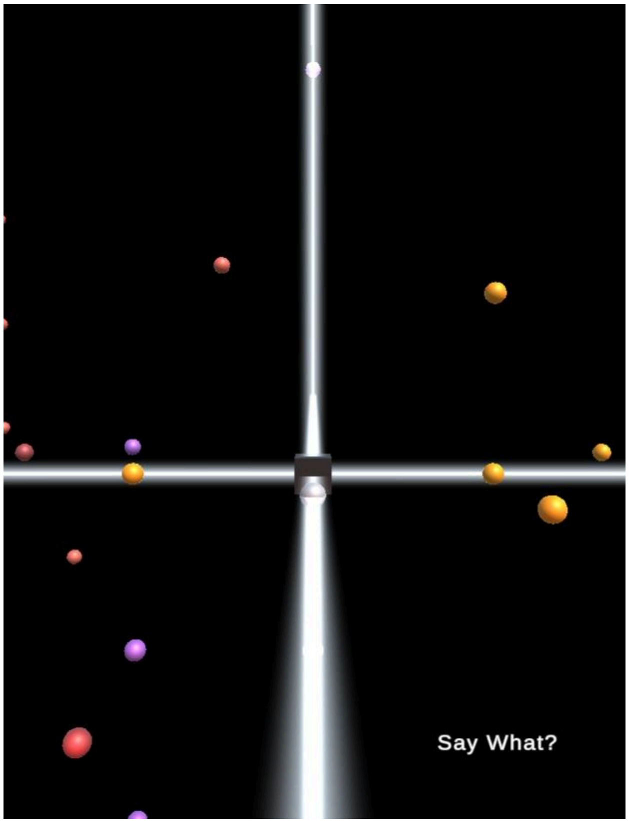

Image 1: Unity3D "Say What?" interactive environment.

[1] Say What? Web application and code available at: https://jfforero.github.io/TheWhatSpace/

The 3D “What” Network — Rhizomatic and Sequential Structures

To further investigate the relational structure of emotional variability, we visualized network topologies using Blender 3D. Two contrasting connectivity models were developed:

- **Rhizomatic network**: Every node (representing a unique utterance) is connected to every other node, forming a fully interconnected, non-hierarchical structure.
- **Sequential network**: Nodes are linked in a closed loop, sequentially traversing the dataset.

Additionally, we explored continuous sequential networks, using cubes placed on curved closed loops to suggest an emotional journey (see Image 2).

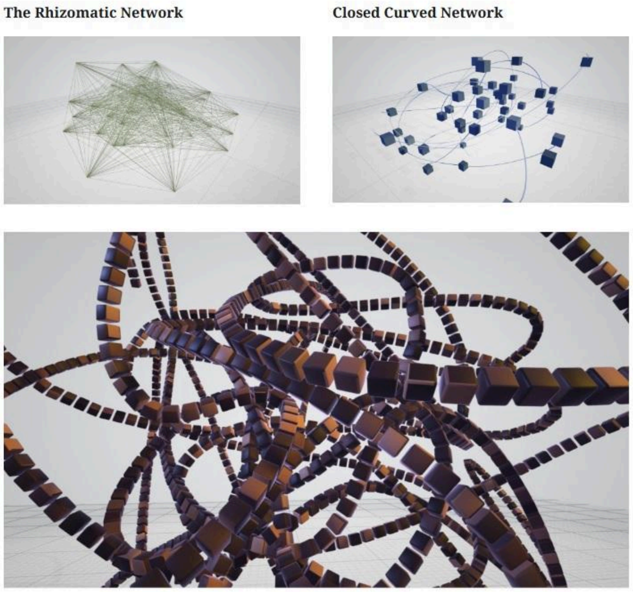

Image 2: Blender “What” Networks

Speech Emotion 2 Texture — Real-Time Emotion-to-Image Translation[2]

We created a Gradio-based web application hosted on Hugging Face Spaces that maps emotional predictions into abstract visual textures. The system pipeline includes:

•Audio input recorded via browser.

•Speech-to-text conversion using OpenAI’s Whisper model (OpenAI, 2022).

•Emotion prediction via pre-trained LSTM-based speech emotion recognition model (Forero, 2023b).

•Emotion-to-image synthesis using DeepAI’s Text-to-Image API[3], prompted with emotion-specific textual descriptions.

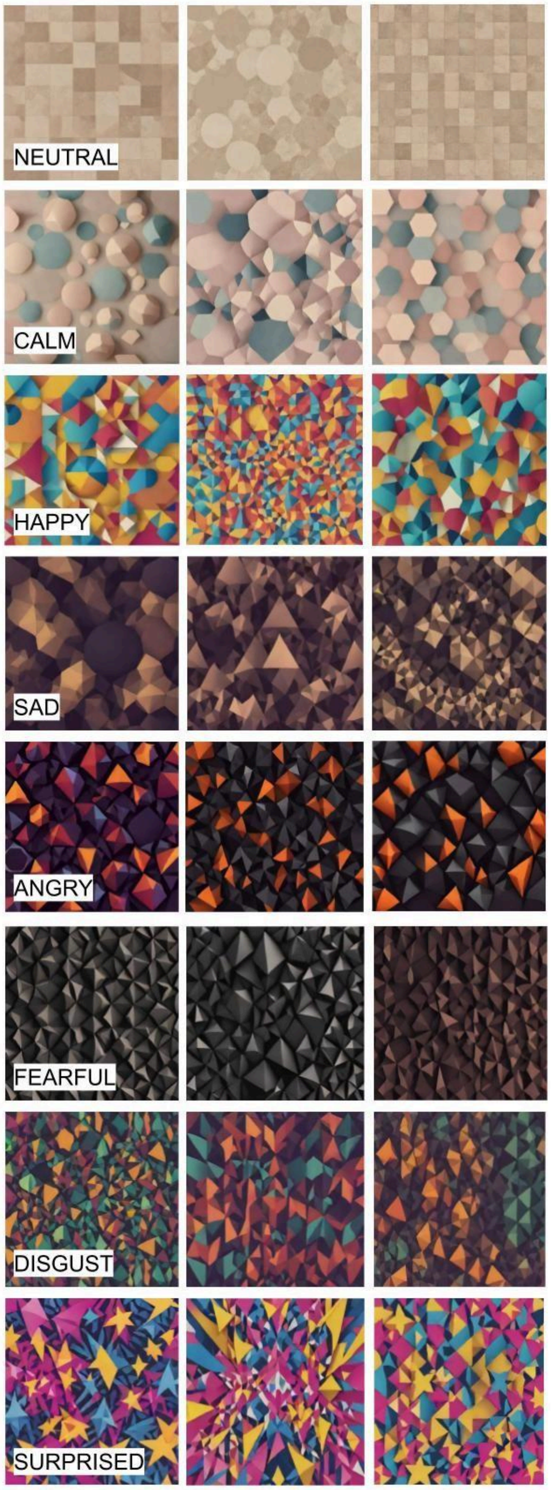

Image 3: Textures created for each emotional prediction

[2] Speech to Texture Hugging Face Space. Available at: https://huggingface.co/spaces/jfforero/LoopArtCritique

[3] DeepAI. (n.d.). Text to Image API. Retrieved from https://deepai.org/machine-learning-model/text2img

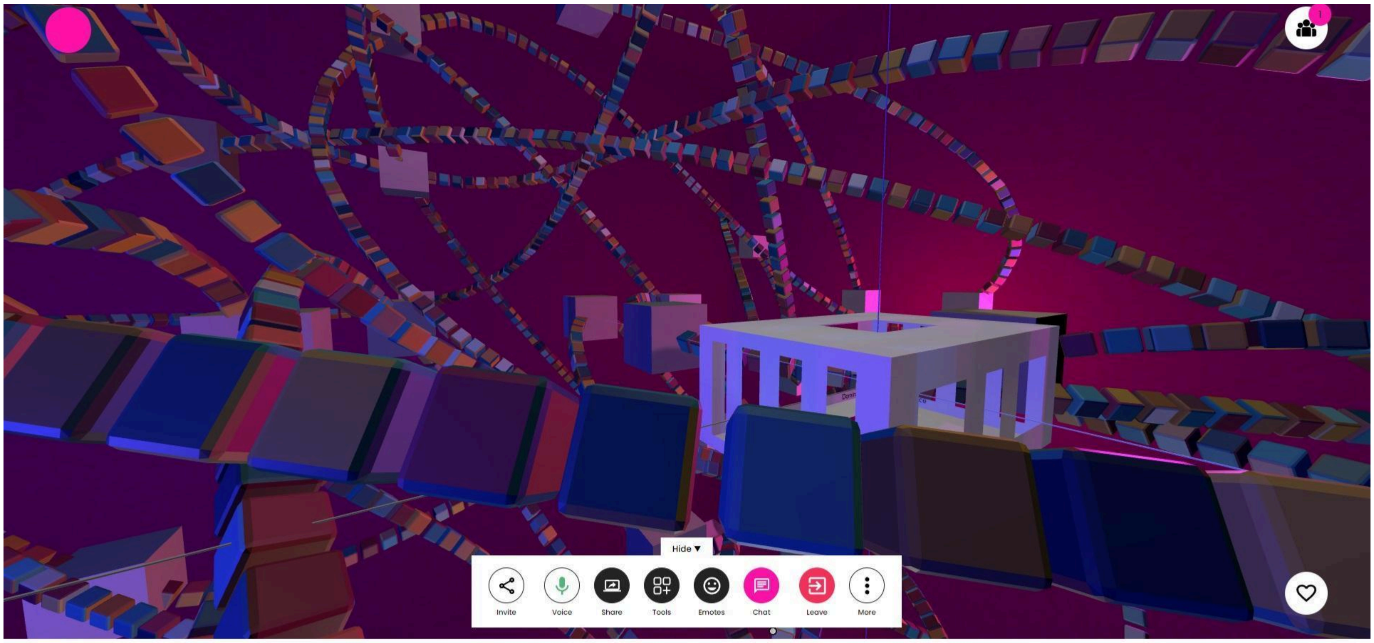

Image 4: The Onland.io “What” verse

The graphical interface allows users to submit audio recordings of the word “what”, encouraging them to experiment with different prosodic variations. In response, the system provides real-time feedback by generating 512px × 512px visual textures that reflect the predicted emotional tone (see Image 3).

The “What” verse[4] — Onland Affectives Virtual Environments

Finally, the project was extended into virtual reality as part of the 13th Loop Art Critique Residency, hosted on the Onland.io metaverse and supported by the MUD Foundation. The installation was built using Mozilla Hubs, an open-source platform for creating shared virtual spaces in WebXR.

In this environment, audio samples of the word “what” were spatialized according to their VAD coordinates and visualized using the sequential network configuration (see Image 4). The experience begins in a silent cuboid space, where users encounter a brief introduction before freely exploring the emotional landscape using a first-person controller in fly mode.

Conclusions and Future Perspectives

This study highlights the expressive richness and emotional ambiguity that can be embedded in minimal linguistic units when analyzed through prosodic variation. By focusing on the single utterance “what,” we demonstrate that emotional interpretation in speech extends beyond semantic content and is critically shaped by acoustic and prosodic features.

The analysis of the IEMOCAP database revealed that “what” occurs across a wide range of emotional categories, with a notable concentration in high-arousal negative emotions such as anger and frustration. However, its presence in lower-arousal and positive emotional contexts further supports its role as a semantically neutral yet affectively versatile utterance.

Through four multimodal visualization strategies—an interactive VAD-space exploration, network-based structures, real-time speech-to-texture mapping, and a virtual reality installation—we translated this emotional variability into spatial and visual forms. These environments offer new frameworks for investigating prosodic emotion and engaging users with affective dimensions of language in interactive settings.

Future work will focus on integrating real-time speech emotion recognition into virtual environments, enabling dynamic updates to textures, spatial mappings, and network structures based on live prosodic input. This direction aims to enhance the development of affect-driven virtual spaces, with implications for affective computing, human-computer interaction, and immersive media research.

Acknowledgements

This work was supported by FCT - Fundação para a Ciência e Tecnologia, I.P. by project reference and DOI identifier: https://doi.org/10.54499/2022.11918.BD

[4] Onland.io Metaverse. Available at: https://verse.loop.onland.io/j4Xy3wf/scaly-hopeful-get-together

References

Aylett, R., & Cavazza, M. (2001). Intelligent virtual environments: A state-of-the-art report. In *Eurographics 2001 Conference* (pp. 1–20).

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4), 335–359.

Camurri, A., Lagerlöf, I., & Volpe, G. (2003). Recognizing emotion from dance movement: Comparison of spectator recognition and automated techniques. *International Journal of Human-Computer Studies*, 59(1–2), 213–225.

Cowie, R., Douglas-Cowie, E., Savvidou, S., & McMahon, E. (2000). FEELTRACE: an instrument for recording perceived emotion in real time. 19-24. Paper presented at Speech and Emotion: *Proceedings of the ISCA workshop*, Newcastle, United Kingdom.

Ekman, P., & Friesen, W. V. (1978). Facial Action Coding System: A Technique for the Measurement of Facial Movement. *Consulting Psychologists Press*.

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3), 572–587.

Forero, J., Bernardes, G., & Mendes, M. (2022). Emotional Machines: Toward Affective Virtual Environments. In *Proceedings of the 30th ACM International Conference on Multimedia* (pp. 1–9).

Forero, J., Bernardes, G., Mendes, M. (2023). Desiring Machines and Affective Virtual Environments. In: *Brooks, A.L. (eds) ArtsIT, Interactivity and Game Creation. ArtsIT 2022. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, vol 479. Springer, Cham.

Jorge Forero, Mónica Mendes, and Gilberto Bernardes. 2023b. En train d'oublier: toward affective virtual environments. In Proceedings of the 20th International Conference on Culture and Computer Science: Code and Materiality (KUI '23). Association for Computing Machinery, New York, NY, USA, Article 2, 1–6.

Misra, V., Schaldenbrand, P., & Oh, J. (2024). Robot Synesthesia: A Sound and Emotion Guided Robot Painter. *arXiv preprint arXiv:2302.04850.*

OpenAI. (2022). Whisper: Robust Speech Recognition via Large-Scale Weak Supervision *[Computer software]*. https://github.com/openai/whisper

Patil, A., Deshmukh, R., & Kulkarni, S. (2023). Emotion-to-Image Translation Using VQGAN+CLIP Framework. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 123–130).

Picard, R. W. (1997). Affective Computing. MIT Press.

Pinilla, A., Garcia, J., Raffe, W., Voigt-Antons, J. N., Spang, R. P., & Möller, S. (2021). Affective visualization in virtual reality: An integrative review. *Frontiers in Virtual Reality,* 2, 630731.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1–2), 227–256.

Sra, M., Garrido-Jurado, S., & Maes, P. (2017). Auris: Integrating Audio and Visual Cues for Immersive Virtual Environments. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 1–12).