The Ironic Machine[s]: Speech Emotion Recognition Methods for Affective Virtual Environment Generation

Jorge Forero Rodríguez

Supervisor
Doutora Mónica Sofia Santos Mendes
Professora Auxiliar
Faculdade de Belas-Artes da Universidade de Lisboa

Second Supervisor
Doutor Gilberto Bernardes de Almeida
Professor Auxiliar
Faculdade de Engenharia da Universidade do Porto

U.PORTO

NOVA
UNIVERSIDADE NOVA
DE LISBOA

U LISBOA | UNIVERSIDADE DE LISBOA

The Ironic Machine[s]: Speech Emotion Recognition Methods for Affective Virtual Environment Generation

Jorge Forero Rodríguez

Doctoral Program in Digital Media

Approved by the Jury:

President: Prof. Doutor João Carlos Pascoal Faria
Referee: Prof. Doutor Nuno António do Nascimento Correia
Referee: Prof. Doutor Hugo Gonçalo Oliveira
Referee: Prof. Doutor Nuno Manuel Robalo Correia
Referee: Prof. Doutor Tiago Barbedo Assis
Referee: Prof. Doutor António Fernando Vasconcelos Cunha Castro Coelho

July 8, 2026

Abstract

The Ironic Machine[s] is a practice-based research project that explores the use of a bimodal speech emotion recognition system for the generation of affective virtual environments and creative computational artifacts. Situated at the intersection of media art, affective computing, and natural language processing, the study investigates how rhetorical figures such as kind-irony and sarcasm emerge through the divergence between semantic and acoustic emotional predictions, and how this interaction can enriches prompt conditioning for the construction of ironic audiovisual spaces and multimodal emotional experiences.

The primary contribution of this thesis is the formulation and empirical evaluation of a method for analyzing illocutionary irony using multimodal speech emotion recognition datasets. Building on sentiment analysis and speech emotion recognition techniques, the research introduces a statistical inference framework that integrates perceptual evaluation with hypothesis-driven analysis to identify prosodic features that significantly differentiate sincere and ironic speech. For posed American English samples, the results demonstrate that divergences between semantic and acoustic emotional predictions can be systematically quantified, supporting the hypothesis that prosodic irony can be understood as a measurable multimodal phenomenon.

A secondary objective investigates whether these findings can be translated into a generative strategy for constructing affective virtual environments. To this end, a three-level prompt engineering framework is proposed, mapping semantic and acoustic emotional predictions onto audiovisual features informed by statistical analyses of musical and visual datasets. Perceptual evaluation indicates that feature-based prompting shows a consistent advantage over high-level emotional labels, although the results do not reach conventional levels of statistical significance.

The investigation is realized through a series of nine technological and artistic iterations, developed as autonomous yet interconnected works. These projects function as experimental platforms for exploring emotional ambiguity, disorientation, and the instability of meaning in computational environments within a techno-poetic narrative.

Collectively, this work proposes both a methodological framework for the analysis of irony and a speculative approach to affective environment generation. It positions divergence between modalities as an operational principle in multimodal systems and demonstrates how practice-based research can serve as a rigorous mode of inquiry into the relationship between human emotion and machine-mediated representation.

Resumo

The Ironic Machine[s] é um projeto de investigação baseado na prática que explora o uso de um sistema bimodal de reconhecimento de emoções na fala para a geração de ambientes virtuais afetivos e artefactos computacionais criativos. Situado na interseção entre a arte dos media, a computação afetiva e o processamento de linguagem natural, o estudo investiga como figuras retóricas como a ironia benevolente e o sarcasmo emergem através da divergência entre as previsões emocionais semânticas e acústicas, e como esta interação pode informar a construção de ambientes virtuais afetivos.

A principal contribuição desta dissertação é a formulação e avaliação empírica de um método para a análise da ironia ilocutória utilizando conjuntos de dados multimodais de reconhecimento de emoções na fala. Com base em técnicas de análise de sentimentos e reconhecimento de emoções na fala, a investigação introduz um enquadramento de inferência estatística que integra avaliação percetiva com análise orientada por hipóteses, de modo a identificar características prosódicas que distinguem significativamente a fala sincera da irónica. Para amostras encenadas de inglês americano, os resultados demonstram que as divergências entre previsões emocionais semânticas e acústicas podem ser sistematicamente quantificadas, apoiando a hipótese de que a ironia prosódica pode ser entendida como um fenómeno multimodal mensurável.

Um objetivo secundário investiga se estes resultados podem ser traduzidos numa estratégia generativa para a construção de ambientes virtuais afetivos. Para esse fim, é proposto um enquadramento de engenharia de prompts em três níveis, mapeando previsões emocionais semânticas e acústicas em características audiovisuais informadas por análises estatísticas de conjuntos de dados musicais e visuais. A avaliação percetiva indica que a utilização de prompts baseados em características apresenta uma vantagem consistente face a rótulos emocionais de alto nível, embora os resultados não atinjam níveis convencionais de significância estatística.

A investigação é realizada através de uma série de nove iterações tecnológicas e artísticas, desenvolvidas como obras autónomas, mas interligadas. Estes projetos funcionam como plataformas experimentais para explorar a ambiguidade emocional, a desorientação e a instabilidade do significado em ambientes computacionais dentro de uma narrativa tecno-poética.

Em conjunto, este trabalho propõe tanto um enquadramento metodológico para a análise da ironia como uma abordagem especulativa à geração de ambientes afetivos. Posiciona a divergência entre modalidades como um princípio operacional em sistemas multimodais e demonstra como a investigação baseada na prática pode constituir um modo rigoroso de inquérito sobre a relação entre a emoção humana e a representação mediada por máquinas.

Preface

This document presents the results of research conducted between 2020 and 2025 in fulfillment of the requirements for the degree of *Doctor in Digital Media, specialization in Technology*, awarded by the Faculty of Engineering of the University of Porto (FEUP). The text brings together a coherent set of theoretical propositions, technical mechanisms, and artistic practices that collectively constitute the operational body of a machine articulated through language and affect[1].

Given my own mode of assemblage, grounded in numerical structures and propositional systems, the initial epistemological approach adopted in this research is predominantly technical. Technique is understood here both as an analytical framework and as a formative structure through which materials are transformed into materialities by means of practice-led research.

Within this framework, each work emerges as a situated possibility that encodes meanings and affects through its specific languages, structures, and cartographies. Considered together, these works document the processes through which the systems presented in this thesis are developed. They also articulate, as a unifying conceptual thread, the irony inherent in the idea of returning "home" when such a home has never existed. In this sense, home is constructed as an imaginary locus, accessible through oral language. Language itself is approached at its Deleuzian limits, informed by concepts such as schizophrenia, desiring-machines, and literary machines.

Accordingly, the material presented to the reader comprises two complementary dimensions. On the one hand, it includes a substantial technical component focused on statistical inference, aimed at identifying and characterizing the properties that differentiate distinct emotional categories. On the other hand, it incorporates a practice-based component developed in a rhizomatic and speculative manner, organized around the divergent—or ironic—concept of a return to the unknown.

1. The complete research repository, including source code and experimental results, is publicly available at https://jfforero.github.io/The-Ironic-Machines/

Figure 1: My daughter Sol portrayed as an infrarealist activist in front of Roberto Bolaño's former residence in Barcelona. *Rim-bot vuelve a casa* is a poetic project based on a BERT model fine-tuned on the literary corpus of Chilean writer Roberto Bolaño. Developed in parallel with this doctoral research, the project informs and modulates its later narrative, incorporating elements from the author's evolving artistic and conceptual framework. The project is available at: https://rimbotvuelveacasa.wordpress.com/.

Acknowledgements

This work was supported by Fundação para a Ciência e a Tecnologia (FCT), Portugal, through national funds, under the doctoral grant awarded to this research (No. 2022.11918.BD).

The investigation presented in this thesis was conducted within the framework of the Doctoral Programme in Digital Media (PMPD) at the University of Porto. It was hosted by the Interactive Technologies Institute (ITI), a research unit of LARSyS – Laboratory of Robotics and Engineering Systems, and at INESC TEC – Institute for Systems and Computer Engineering, Technology and Science.

The author gratefully acknowledges the institutional support, stimulating research environment, and technical resources provided by ITI/LARSyS and INESC TEC, which were essential to the development of this work.

The author expresses his sincere gratitude to his advisor, Mónica Mendes, and to Gilberto Bernardes, for their dedication, guidance, and unwavering commitment throughout this research.

The author also wishes to thank all colleagues, collaborators, and researchers who contributed through insightful discussions, valuable feedback, and continuous support.

This thesis is dedicated to my children Simón, Salvador, and Sol, and to my wife Bárbara. Your love, patience, and resilience made this journey possible.

Jorge Forero Rodríguez
Porto, Portugal

"Sobre la inocente cordillera de la costa
no se sabe más del avioncito
del teniente Bello y su prurito
Por hacer tamaña travesía
cosa que era un acto de heroísmo en sus días
Desafiando nubes y gaviotas
Hacer algo de valor por una cosa
que sólo llevan en el alma los perdidos
más perdidos
El teniente Bello era un perdido
Hermoso estaba de su sueño, herido
Más perdido que el teniente Bello no hay ninguno
desapareció cuando nadie desaparecía
entre Culitrín y Cartagena
cosa que después se hizo corriente un día
Premonitor del sueño y de la pena
de un país por desaparecer
Más perdido que el teniente Bello
Tan hermoso como él"

Mauricio Redolés

Contents

Acknowledgements v

1 Introduction 1
- 1.1 From Intelligent Virtual Environments to Affective Virtual Environments 2
- 1.2 Problem Statement and Research Objectives 10
- 1.3 Research Questions 12
- 1.4 Research Objectives 13
- 1.5 Research Contributions 13
- 1.6 Thesis Structure 16

2 Background and State of the Art 20
- 2.1 Measuring Emotions 21
  - 2.1.1 Categorical Frames 22
  - 2.1.2 Dimensional Frames 24
  - 2.1.3 Reliability, agreement, and quality control 27
- 2.2 Sentiment Analysis 28
  - 2.2.1 Sentiment Analysis Databases 29
  - 2.2.2 Sentiment Analysis Models 30
  - 2.2.3 Evaluation Metrics 32
- 2.3 Speech Emotion Recognition 34
  - 2.3.1 Speech Emotion Recognition Databases 35
  - 2.3.2 Speech Emotion Recognition Models 36
  - 2.3.3 Evaluation Metrics 38
- 2.4 Bimodal Speech Emotion Analysis 41
  - 2.4.1 Irony as Semantic–Prosodic Opposition 41
  - 2.4.2 Acoustic and Prosodic Markers of Irony 43
  - 2.4.3 Cross-Linguistic Prosodic Patterns of Irony 47
- 2.5 Affective Algorithmic Composition 56
  - 2.5.1 Rule-Based Methods 56
  - 2.5.2 Evolutionary Algorithm-Based Systems 58
  - 2.5.3 Neural networks 59
  - 2.5.4 Conditional Neuro-Audio Generation 60
- 2.6 Affective Virtual Environment Generation 63
  - 2.6.1 Rule-Based Methods 63
  - 2.6.2 Image-Based Representations 64
  - 2.6.3 Procedural Encoding 65
  - 2.6.4 Data-Driven Representations 66

3 Methodology **70**
3.1 Emotion Recognition . . . 71
3.1.1 Sentiment Analysis . . . 73
3.1.2 Speech Emotion Recognition . . . 74
3.2 Assessing Irony . . . 76
3.2.1 Perceptual Evaluation . . . 79
3.3 Affective Virtual Environments Generation . . . 80
3.3.1 Music-Based Emotional Profiling . . . 81
3.3.2 Image-Based Emotional Profiling . . . 81
3.4 Integration into the Generator . . . 83
3.4.1 Perceptual Technology Validation . . . 83
3.4.2 The Creative Experience . . . 85
3.5 Ethical Considerations . . . 88

4 Bimodal Speech Emotion Recognition **89**
4.1 Semantic Sentiment Analysis . . . 90
4.1.1 The Rocchio Classifier . . . 91
4.1.2 LSTM Classifier . . . 91
4.1.3 BERT-Based Classifiers . . . 93
4.2 Acoustic Speech Emotion Recognition . . . 96
4.2.1 MLP Classifier . . . 96
4.2.2 CNN Classifier . . . 99
4.2.3 CNN+LSTM Classifier . . . 101

5 Analyzing Verbal Irony through Prosodic–Semantic Divergence **104**
5.1 Data and Feature Selection . . . 105
5.2 Statistical Analysis . . . 108
5.3 Perceptual Evaluation . . . 112

6 Affective Virtual Environment Generation **116**
6.1 Affective Music Generation . . . 117
6.1.1 Data and Feature Selection . . . 117
6.1.2 Statistical Analysis . . . 118
6.2 Affective Visual Generation . . . 121
6.2.1 Data and Feature Selection . . . 122
6.2.2 Statistical Analysis . . . 122
6.3 A Three-Level Prompt Engineering Framework for Affective Environment Generation . . . 126
6.3.1 Perceptual Evaluation . . . 126

7 Creative Practice **132**
7.1 Iteration 1: What Stirs Within . . . 137
7.2 Iteration 2: A Máquina Sensível . . . 139
7.3 The Emotional Machine[s] . . . 142
7.3.1 Iteration 3: Abandoned Memories . . . 147
7.3.2 Iteration 4: Veintitantos Vientos . . . 148
7.3.3 Iteration 5: En train d'oublier . . . 151
7.4 Iteration 6: Terra Australis Ignota . . . 157
7.5 Iteration 7: The What Space . . . 164

7.6 Iteration 8: The Ironic Machine[s] 168
7.7 Iteration 9: Bello 170

8 Conclusion **177**
8.1 Dissertation Contributions 178
8.2 Research Questions Addressed 180
8.3 Limitations and Future Directions 182

References **184**

A Documentation and Repository Structure **207**
A.1 Technical Situation: Scientific Methodology 208
A.2 Practical Situation: Iterative Development 208
A.3 Visual Documentation and Artifacts 209
A.4 Poe++ 210

List of Figures

1 My daughter Sol portrayed as an infrarealist activist in front of Roberto Bolaño's former residence in Barcelona. *Rim-bot vuelve a casa* is a poetic project based on a BERT model fine-tuned on the literary corpus of Chilean writer Roberto Bolaño. Developed in parallel with this doctoral research, the project informs and modulates its later narrative, incorporating elements from the author's evolving artistic and conceptual framework. The project is available at: https://rimbotvuelveacasa.wordpress.com/. iv

1.1 The Cave Automatic Virtual Environment (CAVE) at the Electronic Visualization Laboratory (EVL), University of Illinois at Chicago. Photograph by Dave Pape (self-timed capture), Public Domain, via Wikimedia Commons: https://commons.wikimedia.org/w/index.php?curid=868395. 4

1.2 *Hidden Worlds of Noise and Voice* (2002), an augmented-reality installation by Golan Levin and Zachary Lieberman, developed at the Ars Electronica Futurelab and exhibited at the Ars Electronica Museum, Linz. The system employed custom optical see-through headsets with position tracking and embedded microphones to visualize users' vocal utterances as animated three-dimensional forms appearing to emanate from the speaker's mouth. While these volumetric graphics were visible only through the head-mounted displays, projected "shadows" of the virtual objects on the table surface allowed non-wearing spectators to observe the interaction. Source: https://www.flong.com/archive/projects/hwnv/index.html. . . 6

1.3 Technical configuration diagram of *RE:MARK* and *Hidden Worlds of Noise and Voice*, installed at the Ars Electronica Center (September 2002). The schematic (top) illustrates the shared computational infrastructure supporting both installations. The lower image shows *RE:MARK* in situ at the Ars Electronica Center. Source: Levin, G., Lieberman, Z., AEC Futurelab (2002). Available at https://personales.upv.es/moimacar/pages/catalog/03/image/golan/system.pdf. . . 7

1.4 Overview of the *Auris* pipeline for generating affective virtual environments from music. Song audio is processed through MFCC-based mood classification, while song lyrics undergo noun-phrase extraction. The resulting affective (mood) and semantic (phrase) representations jointly condition a Mood-Conditional PixelCNN image generator. The synthesized images are transformed into textures and mapped onto 3D assets, producing a procedurally composed VR scene. Adapted from Sra et al. 2017. 9

1.5 Examples of VR environments generated by the *Auris* system. The scenes illustrate textured 3D objects whose materials are derived from DeepDream-enhanced images conditioned on musical mood and lyrical content. The resulting immersive environments visually encode affective characteristics inferred from the input songs. Adapted from Sra et al. 2017. 9

2.1 (a) Russell's Circumplex Model, illustrating the distribution of affective states along the dimensions of valence (pleasure) and arousal (activation). (b) Plutchik's Wheel of Emotions, depicting the intensity, relationships, and polarities of primary and secondary emotional categories. 26
2.2 The Self-Assessment Manikin (SAM) interface (Bradley and Lang 1994) for measuring affect along the dimensions of valence (pleasure), arousal (activation), and dominance (control). The pictorial format enables rapid, language-independent emotional self-report. Image adapted from Sainz-de-Baranda Andujar et al. (2022). 27
2.3 Rule-based mapping between musical parameters and emotional expression in the valence–arousal space. The diagram illustrates how compositional features—including tempo, mode, harmony, pitch register, articulation, and timbre—are associated with discrete affective quadrants (e.g., active–positive, passive–negative), enabling systematic transformation of musical material to induce target emotional states. Adapted from Livingstone et al. (2010). 57
2.4 Overview of semantic music generation in *MusicLM*. The architecture integrates neural audio tokenization (SoundStream), self-supervised acoustic representation learning (w2v-BERT), and joint audio–text embedding via contrastive learning (MuLan) to align natural language descriptions with learned musical structure. This multimodal representational hierarchy enables coherent high-fidelity audio synthesis directly from rich natural language prompts, as illustrated by the mapping of descriptive captions to synthesized musical outputs. Adapted from Agostinelli et al. (2023). 61
2.5 Semantic-to-spatial generation pipeline in Hunyuan World. A natural language prompt (left) is first mapped into a latent semantic representation, which is used to synthesize a globally coherent panoramic scene (center). From this representation, the model generates consistent multi-view renderings (right), preserving object geometry, lighting, and spatial relationships across viewpoints. By learning spatiotemporal correlations from large-scale video data, the system enables persistent, navigable environments that approximate interactive world simulation from textual input. 69

3.1 Architecture of the Affective Virtual Environment (AVE) System. The workflow illustrates the transformation of input speech into an immersive virtual environment. The process involves two primary mechanisms: an emotion recognition system, using acoustic and semantic models to classify affect, and a virtual environment generation system that maps these classifications to corresponding audio-visual attributes for VR rendering. 70
3.2 Detailed Architecture of the Affective Virtual Environment System. This diagram illustrates the dual-stream processing for emotional mapping within the AVE System. 72
3.3 Experimental pipeline for ironic sample filtering and analysis. This architecture details the multimodal processing of emotional nuances, specifically focusing on the distinction between sincere anger vs. kind-ironic speech, and sincere happy vs. sarcastic speech. 77
3.4 PODONOS web interface for perceptual audio analysis, used to upload and process speech samples. 79
3.5 Mapping strategy for affective asset generation. The diagram depicts the transformation of multimodal emotional predictions into descriptive prompts for the synchronized synthesis of virtual environment components. 80

3.6 The web platform interface used for the perceptual evaluation of generated environments. Participants rated the alignment of audiovisual stimuli on a five-point Likert scale ranging from *Perfectly Matched* to *Completely Mismatched*. 84
3.7 Rhizomatic network of the nine creative artifacts developed throughout the research. Each node represents an individual project, annotated with its primary technological or methodological focus. The diagram visualizes the non-linear, practice-based research process, in which successive artifacts emerge through recursive experimentation across domains such as speech analysis, machine learning, and audiovisual generation, collectively contributing to the development of the Affective Virtual Environment (AVE) framework. 86
3.8 Project landing page (www.emotional-machines.com), developed and maintained during the research process. The website serves as a centralized resource for technical specifications and implementation details related to the creation of affective virtual environments. 87

4.1 Confusion matrices for the Rocchio Classifier applied to DynaSent Round 1. Labels 0, 1, and 2 represent negative, neutral, and positive sentiment, respectively. The matrices illustrate that while the Cleaned variant (left) achieves the best overall balance, the Lemmatized version (right) exhibits superior recall for the neutral class at the expense of negative class sensitivity. 92
4.2 Confusion matrices for the LSTM model using randomly initialized embeddings (Without GloVe). The results illustrate the classification asymmetry noted in Table 4.3, characterized by high negative-class recall but frequent misclassification of neutral and positive instances into the negative category. 93
4.3 Confusion matrices for the BERT (left) and BERT++ (right) classifiers. Compared to previous models, the BERT-based architectures demonstrate significantly more balanced performance across negative, neutral, and positive classes, with the BERT++ variant showing the most robust diagonal alignment. 95
4.4 Confusion matrix for the MLP classifier trained on eGeMAPS acoustic features. . 97
4.5 Confusion matrix for the MLP classifier trained on MFCC features. 98
4.6 Confusion matrix for the 1D CNN classifier trained on eGeMAPS features (original dataset). 100
4.7 Confusion matrix for the 1D CNN classifier trained on augmented eGeMAPS features. 101

5.1 Overview of the IEMOCAP Multimodal Database. (Top) Representative text transcriptions paired with their corresponding emotional labels. (Bottom) Distribution analysis showing the total number of samples categorized per emotion across the dataset. 106
5.2 Emotion distribution for the lexical token "What?": A frequency analysis illustrating the prosodic variability of a single word across the IEMOCAP dataset. 107
5.3 Circular visualization of PCA loadings for Kind Irony vs. Sincere (Anger): feature contributions to principal components for male (left) and female (right) speakers. 110
5.4 Circular visualization of PCA loadings for Sarcasm vs. Sincere (Happy): feature contributions to principal components for male (left) and female (right) speakers. 112
5.5 Mean irony ratings across affective speech categories. Error bars represent 95% confidence intervals. 113

6.1 Distribution of musical excerpts across genres in the 4Q Audio Emotion Database. Pop/Rock represents the largest category, followed by Jazz and Electronic music. This distribution motivated the selection of the Pop/Rock subset for the affective feature analysis. 117
6.2 Frequency of top mood tags in Pop/Rock tracks. High-arousal negative descriptors (*angry*, *harsh*, *hostile*) co-occur with positive descriptors (*happy*, *joyous*), confirming the suitability of this subset for valence-arousal analysis. 118
6.3 Box plots showing the distribution of discriminative musical features across the four emotional quadrants. Variations in spectral, harmonic, rhythmic, and dynamic descriptors confirm differential affective profiles and are consistent with the arousal–valence structure of Russell's circumplex model. Features marked with an asterisk (*) in Table 6.1 survived FDR correction. 119
6.4 **3D Normalized HSV Color Histograms (512 Bins) across Emotional Categories.** Each plot visualizes the global color distribution using an $8 \times 8 \times 8$ binning scheme. The **spatial position and color** of each sphere correspond to its specific bin in the Hue-Saturation-Value coordinate system, while the **sphere radius** denotes the relative contribution (normalized frequency) of that bin to the emotional profile. Consistent with the statistical analysis, *Fear* and *Sadness* show pronounced activation (larger radii) in the low-Value and low-Saturation regions. Conversely, *Joy* exhibits a shift toward higher-Value bins and more vivid chromatic clusters, confirming its status as the most visually bright and saturated category ($\eta^2 = 0.109$). 124
6.5 Visualization of the 64 generated environments used in the perceptual evaluation. The grid shows eight scenes for each audiovisual condition across two prompting strategies (high-level and mid-level). Conditions correspond to the proposed irony framework: Kind Irony (negative image + positive audio), Sarcasm (positive image + negative audio), Sincere Happy (positive image + positive audio), and Sincere Anger (negative image + negative audio). 128

7.1 Public installations of *Terra Australis Ignota*. Top: immersive projection mapping inside the *Cattedrale dell'Immagine* during the *XR Festival Florence* (2024), where cartographic imagery and generated creatures were projected across the cathedral's architectural surfaces. Middle and Bottom: extended multi-surface projection installation presented at the *Noite Europeia dos Investigadores* at the National Museum of Natural History and Science, Lisbon, where the work was adapted to a gallery environment using panoramic projections across multiple walls. 134
7.2 Public exhibition flyer for the 13th Ambience on Onland.io. The What space. . . 135
7.3 Architectural projection mapping of *Bello* at the Igreja da Sé, Faro, Portugal, presented during the *Algarve Design Meeting*. 136
7.4 Project installations at M.ou.co (top) and ACM (bottom), showcasing both screen-projection and immersive VR presentation modes. 137
7.5 System architecture for real-time affective audio synthesis: (Left) Web-based front-end interface capturing speech transcription and user-defined emotional parameters; (Right) Corresponding Pure Data (Pd) patch performing dynamic audio modulation via the WebPd framework. 138
7.6 Voice-command interaction workflow in *A Máquina Sensível*. The user issues spoken commands that are transcribed through the speech recognition module and mapped to scene operations, enabling the real-time creation and transformation of virtual objects (e.g., spheres, humanoid figures, and color modifications). 139

7.7 System architecture of *A Máquina Sensível*. (Top) Command-recognition logic showing the hierarchical organization of methods, categories, and keywords used to interpret spoken instructions. (Center) Transformation and attribute operations available for virtual objects, including colorization, translation, rotation, and scaling. (Bottom) Graphical Help interface presenting the available voice commands and interaction instructions for the user. 140
7.8 Screenshot of the Affective Virtual Environment (AVE) system developed between 2021 and 2022. The image shows the result of a user wearing a VR headset speaking the phrase "we were young and we were crazy". The generated image derived from this text prompt is subsequently used as a style source to transform a second image (an equirectangular panorama) selected according to the sentiment analysis of the spoken input. The interface also displays the GUI with two VR buttons ("SPEAK" and "GENERATE") and the real-time transcription of the user's speech. 142
7.9 Flowchart for affective texture synthesis: The workflow integrates speech-to-text transcription with sentiment analysis to compose generative prompts, which are then used to produce style images for final equirectangular target image transformation. 143
7.10 Interface overview of a synthesized Affective Virtual Environment (AVE). The screenshot displays the real-time mapping of speech-derived sentiment onto a 3D object's texture, illustrating the convergence of text, target imagery, and generated emotional profiles. 144
7.11 Updated Affective Virtual Environment user interface. The system demonstrates the real-time integration of speech-to-text transcription, sentiment analysis, and the resulting generative texture synthesis applied to the immersive 3D scene. . . . 144
7.12 Mapping of 3D spatial data for immersive environments: (Top) Six-face Cube Map representation of a virtual scene; (Bottom) The resulting Equirectangular Panorama used as the target image for affective style transfer. 146
7.13 Conchord Extended: Graphical user interface description for Music generation using the tonal interval space in Pure data 147
7.14 Affective texture generation via Neural Style Transfer. The process demonstrates the transformation of a base equirectangular panorama of "Valle de la Muerte" (left) into a stylized texture guided by the emotional prompt "desolate place" (right), resulting in a visually congruent environment for low-valence affective states. 148
7.15 The *Emotional Machines* interactive interface. The GUI facilitates the input of geographic coordinates alongside real-time voice recording to synthesize immersive equirectangular panoramas through affective image diffusion and text-guided stylization. 152
7.16 The visual poem *Veinte y tantos Vientos*: A collection of 24 equirectangular panoramas generated using the *Emotional Machines* system. Each image represents a distinct geospatial coordinate within the Campo 24 de Agosto area in Porto, Portugal, synthesized through text-guided affective diffusion. 153

7.17 Binary interface from the web-based artwork *En train d'oublier*. Each panorama is split by an interactive vertical slider that allows viewers to reveal two simultaneous states of a generated environment. On the left, the landscape corresponds to the abandoned railway site at the Salar de Uyuni, while on the right the same environment is algorithmically transformed into speculative or impossible futures. The conjugated forms of the verb *être* ("Je suis," "Nous sommes") are paired with the phrase *en train d'oublier*, framing identity as an ongoing process of becoming and forgetting. 154
7.18 The visual poem *En train d'oublier*: a collection of equirectangular panoramas generated using the *Emotional Machines* system. Each image is a distinct transformation of the same site through text-guided affective diffusion. 156
7.19 Left: Example of Emotional Maps generation based on the fusion of emotional speech and curated cartographic imagery. Right: Example of Emotional Monsters generation, where emotion and language drive the direct creation of synthetic creatures. 159
7.20 Folder containing image files created with the Monster generator system. 160
7.21 Maps are created by searching a location in Google Maps and modify by the text transcribed and the emotional predictions 161
7.22 Distribution of "what" utterances in VAD space. Each sphere corresponds to a sample listed in the accompanying table, which details its emotional label and VAD values. The dispersion illustrates how identical lexical content gives rise to distinct affective states through prosodic variation. 164
7.23 Network-based representations of the "What" space. (Top-left) Rhizomatic network showing a fully interconnected, non-hierarchical structure of utterances. (Top-right) Closed sequential network linking samples in a continuous loop. (Bottom) Continuous curved trajectories composed of discrete nodes, suggesting an affective journey through prosodic variation. 165
7.24 Emotion-conditioned texture generation from vocal input. Each row corresponds to a predicted emotional category (e.g., neutral, calm, happy, surprised, sad, fearful, angry, disgust), with multiple samples illustrating how predefined prompts translate prosodic features into variations in color, geometry, and illumination. . . 166
7.25 Immersive VR environment of the What Space on Onland.io. The scene presents a dense network of interconnected nodes arranged along looping trajectories, where each element corresponds to a recorded utterance of the word "what". Spatialized structures and chromatic variations reflect affective differences derived from prosodic analysis, allowing users to navigate an emotionally distributed topology through first-person exploration. 167
7.26 Examples of AI-generated texture images for two contrasting emotional categories—*Joy* (top) and *Anger* (bottom)—across three levels of prompt specificity. Each row corresponds to a different prompting strategy derived from the framework in Section 6.3: low-level (feature-based parametric descriptions), mid-level (semantic descriptors grounded in statistically significant audiovisual features), and high-level (abstract emotional labels). The results illustrate how increasing levels of abstraction influence visual appearance, with low-level prompts producing more controlled and homogeneous textures, and high-level prompts yielding more diverse and stylistically varied outputs. 169

7.27 Interface of the *Ironic Machine* implemented in p5.js, displaying the four audio-visual configurations derived from cross-modal emotional combinations. Each panel represents a distinct condition: *Sincere Happiness* (top-left), *Kind Irony* (top-right), *Sarcasm* (bottom-left), and *Sincere Anger* (bottom-right). The system allows real-time selection of visual and musical emotional profiles, enabling the exploration of alignment and divergence between modalities as a mechanism for producing ironic affective environments. 170
7.28 Excerpt from the poem *El Teniente Bello* by Mauricio Redolés, originally featured in the album *Bello Barrio* (1987). The poem reinterprets the historical disappearance of Teniente Bello through a political and poetic lens, linking themes of memory, loss, and national identity. In this work, the text is incorporated as a semantic and narrative substrate within the generative system, connecting literary discourse with speech-driven affective analysis and audiovisual synthesis. 173
7.29 Visual representation inspired by the figure of Teniente Bello, integrating poetic structure and generative imagery. 176

A.1 Overview of the GitHub repository structure and technical situations. 207
A.2 Visual documentation of the iterative process, showcasing screenshots from the primary Affective Virtual Environments (AVE). 209
A.3 Vous êtes fou, monsieur Rim-bot? 211

List of Tables

1.1 Operational Tasks and Corresponding Research Questions 13

2.1 Taxonomies of primary emotion sets: a summary of foundational and contemporary classification models in affective science. 23
2.2 Evolution of multidimensional emotion models: A comparative overview of foundational axes used to map affective states. 25
2.3 Summary representative sentiment analysis datasets and their characteristics. . . 30
2.4 Sentiment classification performance across model paradigms. 33
2.5 Representative Speech Emotion Recognition datasets and their characteristics. . . 36
2.6 Speech Emotion Recognition performance across model paradigms. 39
2.7 Acoustic analysis of dry and dripping ironic versus non-ironic utterances (Bryant and Fox Tree 2005) 44
2.8 Summary of acoustic cues distinguishing sarcasm from other attitudes (Cheang and Pell 2008) 44
2.9 Sarcasm classification results on the MUStARD dataset (Castro, Hazarika, Poria, et al. 2019) 46
2.10 Production findings and perception recognition rates for sarcasm and kind irony (B. Braun and Anna Schmiedel 2018) 48
2.11 Prosodic Irony Studies 50

3.1 Semantic-Acoustic opposition in ironic configurations (Kind Irony and Sarcasm) vs. the sincere expressions in literal speech. 76
3.2 Prosodic features extracted for statistical analysis, including descriptions and measurement units. 78
3.3 Audio feature set extracted with the Essentia library, organized by musical dimension. These descriptors capture pitch prominence, harmonic structure, rhythmic activity, loudness, and spectral brightness, properties widely associated with emotional expression in music. 81
3.4 Visual feature set extracted from the Emotion6 dataset, organized by visual dimension. These descriptors capture chromatic distribution, textural complexity, compositional balance, and structural properties widely associated with emotional expression in images. 82
3.5 Cross-Modal Affective Mapping Logic: Combinatorial rules for synthesizing ironic and sincere audiovisual spaces based on visual and musical emotional profiles. . 83

4.1 Representative examples of original, cleaned, and lemmatized text from the DynaSent dataset, illustrating the cumulative transformations applied at each preprocessing stage. 90

4.2 Test-set evaluation of Rocchio classifiers on DynaSent Round 1. Class 0 = negative, 1 = neutral, 2 = positive. All figures report precision (P), recall (R), and F1-score (F1) per class, together with overall accuracy (Acc.). 91
4.3 Test-set evaluation of LSTM models on DynaSent Round 1. Class 0 = negative, 1 = neutral, 2 = positive. Results are averaged across three independent runs with distinct random seeds. 92
4.4 Test-set evaluation of BERT-based classifiers on DynaSent Round 1. Class 0 = negative, 1 = neutral, 2 = positive. 94
4.5 Acoustic feature set extracted using openSMILE eGeMAPSv02, organized by paralinguistic domain. 96
4.6 Classification report for MLP on RAVDESS (eGeMAPS features). 97
4.7 Classification report for MLP on RAVDESS (MFCC features). 98
4.8 Classification report for CNN on RAVDESS (eGeMAPS features, original dataset). 99
4.9 Classification report for CNN on RAVDESS (eGeMAPS features, augmented dataset). 101
4.10 Classification report for CNN+LSTM on RAVDESS (eGeMAPS features, augmented dataset). 102

5.1 Bimodal configuration of semantic sentiment and prosodic emotion used to identify ironic and sincere speech categories. 108
5.2 Descriptive summary of selected prosodic features and their corresponding units of measurement. 108
5.3 Comparative Statistical Analysis of Prosodic Features. This table presents the mean and standard deviation for vocal descriptors across gender, identifying statistically significant differences ($p < 0.05$) between sincere and ironic expressions to validate the experimental filtering pipeline. 109
5.4 Statistical Analysis of Vocal Features by Gender and Expression Type (Sincere vs. Sarcasm) 111
5.5 Descriptive statistics of perceived irony ratings (stimulus-level means) by category. 113

6.1 Comparison of musical features across emotional quadrants (mean $\pm$ standard deviation). Asterisks (*) denote features surviving Benjamini–Hochberg FDR correction ($\alpha = 0.05$). Non-significant features are included for completeness. . . . 120
6.2 Comparison of selected visual features across emotional categories (mean $\pm$ standard deviation). Asterisks (*) denote features surviving Benjamini–Hochberg FDR correction ($\alpha = 0.05$). Values for brightness are in pixel intensity units (0–255); hue and saturation are in HSV scale (0–255); all other features are unitless proportions or normalized scores. 122
6.3 Three-Level Prompting Framework for Affective Environment Generation 127
6.4 Average perceived audiovisual mismatch scores for each generated environment. Lower values indicate stronger perceived alignment between modalities. 129
6.5 Average mismatch ratings grouped by audiovisual emotional configuration. . . . 129
6.6 Average perceived audiovisual mismatch by prompting strategy. 130

7.1 Emotion-conditioned prompts used to guide text-to-image generation. Each prompt encodes visual attributes such as color, illumination, and geometric structure associated with the predicted emotional state. 167
7.2 Audiovisual profile combinations used to generate the Ironic Machine[s]. 168

7.3 Sentiment-conditioned prompt generation for equirectangular image synthesis. Visual attributes are modulated according to sentiment polarity derived from transcribed speech. Note that the [**transcribed text**] is insert in the prompt 171
7.4 Emotion-conditioned prompt generation for music synthesis using MusicGen. Acoustic emotion predictions modulate temporal, spectral, and harmonic properties of the generated audio. 172

A.1 Summary of Technical Implementation Notebooks 208

List of Acronyms

4Q	Four-Quadrant (Russell Circumplex Model)
AFINN	AFINN Sentiment Lexicon
AI	Artificial Intelligence
ANOVA	Analysis of Variance
ANVIL	A Notation for Video and Language
API	Application Programming Interface
AR	Augmented Reality
ASR	Automatic Speech Recognition
AVE	Affective Virtual Environments
BERT	Bidirectional Encoder Representations from Transformers
BPM	Beats Per Minute
CASE	Computer-Aided Software Engineering
CI	Confidence Interval
CNN	Convolutional Neural Network
CORBA	Common Object Request Broker Architecture
CAVE	Cave Automatic Virtual Environment
DDSP	Differentiable Digital Signal Processing
EDA	Electrodermal Activity
EEG	Electroencephalography
FDR	False Discovery Rate
FMA	Free Music Archive
GUI	Graphical User Interface
HCI	Human–Computer Interaction
HMD	Head-Mounted Display
HNR	Harmonics-to-Noise Ratio
HOG	Histogram of Oriented Gradients
HPCP	Harmonic Pitch Class Profile
HSV	Hue–Saturation–Value
IEMOCAP	Interactive Emotional Dyadic Motion Capture
IVE	Intelligent Virtual Environment
IVELL	Intelligent Virtual Environment for Language Learning
JS	JavaScript
KMeans	K-Means Clustering
LBP	Local Binary Patterns
LLM	Large Language Model
LSTM	Long Short-Term Memory Network
MFCC	Mel-Frequency Cepstral Coefficients
MLP	Multilayer Perceptron

MR	Mixed Reality
NLP	Natural Language Processing
OOV	Out-Of-Vocabulary
PCA	Principal Component Analysis
Pd	Pure Data
ReLU	Rectified Linear Unit
RGB	Red–Green–Blue
ROI	Region of Interest
RAVDESS	Ryerson Audio-Visual Database of Emotional Speech and Song
SEM	Standard Error of the Mean
SER	Speech Emotion Recognition
SOAR	State, Operator, And Result
TF-IDF	Term Frequency–Inverse Document Frequency
TIM	The Ironic Machine (corpus)
VAD	Valence–Arousal–Dominance
VE	Virtual Environment
VIEW	Virtual Interface Environment Workstation
VR	Virtual Reality
WebXR	Web Extended Reality
WebPD	WebPD (Pure Data for the Web)
WWW	World Wide Web

Chapter 1

Introduction

This work is an emotional contradiction

Emotions are often described as brief, relatively intense affective episodes that arise when an individual appraises a situation as significant (Arnold 1960; Lazarus 1991). They are typically understood as part of the broader domain of affect, which also includes moods and enduring dispositions (Juslin and Sloboda 2010). A further distinction can be made between felt, expressed, and perceived emotions, highlighting that subjective experience, outward display, and external attribution do not necessarily coincide (Scherer 2004; Barrett, Satpute, and Bliss-Moreau 2019).

Emotions are deeply intertwined with language: semantic descriptors provide a conceptual structure through which affective experiences are categorized, communicated, and socially negotiated (Shaver et al. 1987; Barrett, Satpute, and Bliss-Moreau 2019). Through language, emotions become not only internal states, but shared constructs, shaped by cultural conventions and communicative practices.

Emotions have increasingly been incorporated into computational systems and interactive environments. In affective computing, Picard (1997) argued that emotions are not opposed to rationality, but constitute an essential component of intelligent behavior, enabling machines to recognize, interpret, and respond to human affective states. Within virtual environments, Aylett and Luck (2000) described Intelligent Virtual Environments (IVE) as systems in which narrative, character behavior, and interaction are dynamically structured through AI-driven mechanisms or agents.[1]

Building on this paradigm, Affective Virtual Environments (AVE) extend the concept of intelligent environments by integrating automatic emotion recognition with generative audiovisual systems, enabling environments that adapt dynamically to users' affective states (Pinilla et al. 2021). Emotion recognition in such systems is commonly performed through multiple modalities, including facial and body analysis (Ekman and Friesen 1971; Li and Deng 2020), acoustic and prosodic

1. An agent can be defined as an entity that perceives its environment through sensors and acts upon that environment through actuators (Russell and Norvig 2010).

speech analysis (Scherer 2003; Eyben, Wöllmer, and Schuller 2010), natural language processing of semantic content (Pang and Lee 2008; Devlin et al. 2019), and physiological signals such as electroencephalography (EEG), heart rate (HR), respiration, and galvanic skin response (GSR) (Koelstra et al. 2012; Egger, Ley, and Hanke 2019; Hofmann et al. 2021; Zhang et al. 2017).

Among these modalities, speech occupies a uniquely expressive and analytically rich position. It conveys complementary layers of affective information within a single signal (Mehrabian and Ferris 1967). In contexts such as virtual reality, where facial cues may be occluded by head-mounted displays, speech becomes a primary channel for affect inference. Crucially, speech comprises two analytically distinct yet dynamically interacting channels: a semantic channel encoding propositional meaning, and an acoustic channel encoding prosodic features such as pitch, intensity, and temporal dynamics (Scherer 2003; Busso et al. 2008).

When considered jointly, these channels do not simply complement each other, but may also diverge. In this thesis, we argue that such divergence constitutes a fundamental mechanism for meaning-making in speech. In particular, we conceptualize verbal irony as a structured mismatch between semantic polarity (what is said) and prosodic valence (how it is said).

Rather than treating irony as a discrete emotional attitude, this work models it as a relational configuration emerging from cross-modal incongruence. This perspective enables the systematic detection, analysis, and generation of ironic expressions within computational systems.

This framework is further extended to the generation of Affective Virtual Environments (AVE), in which bimodal emotional representations are mapped onto audiovisual parameters through structured prompt engineering strategies. The research is materialized through a series of practice-based artistic iterations, which function as experimental platforms for testing and extending the proposed framework. These works enable the exploration of emotional ambiguity, multimodal divergence, and the instability of meaning within computational environments.

1.1 From Intelligent Virtual Environments to Affective Virtual Environments

Intelligent Virtual Environments (IVEs) emerged from the convergence of real-time 3D virtual environments and Artificial Intelligence techniques. In their state-of-the-art report, Aylett and Cavazza (2001) define Intelligent Virtual Environments as systems in which AI mechanisms are embedded within the architecture of the virtual world itself, enabling adaptive behavior, problem-solving capabilities, and autonomous virtual agents. Within this framework, interaction modalities, particularly natural language and speech, became central mechanisms through which users could communicate with intelligent entities and influence virtual world dynamics. The integration of speech into virtual environments has evolved over the past decades, progressing from early user-dependent, command-based systems to multimodal, user-independent systems powered by deep neural automatic speech recognition and large language models (Reddy 1976; Bolt 1980; Radford, Kim, et al. 2023). Across this period, speech-based interaction has been deployed in multiple

application domains, including aerospace simulation (Furness 1986; Fisher et al. 1989), education and training (Johnson, Rickel, and Lester 1998; Johnson, Marsella, and Vilhjálmsson 2004), healthcare (Ochs and Pelachaud 2017; Rizzo et al. 2014), entertainment (Ubisoft 2017; Bethesda Game Studios 2017), and artistic installations (Levin and Lieberman 2004; Sra et al. 2017; Ryu 2023).

One of the earliest applications of speech recognition in virtual reality appeared in flight simulators during the early 1980s. Thomas Furness's *Super Cockpit* system enabled pilots to issue spoken commands alongside manual and gaze-based controls (Furness 1986). Building on this development, British Aerospace introduced the *Virtual Cockpit* in 1987, integrating a head-mounted display with automatic speech recognition to support hands-free pilot interaction (British Aerospace 1987). Further advances emerged at NASA's Ames Research Center with the development of the Virtual Interface Environment Workstation (VIEW) in 1989. VIEW combined real-time computer graphics, spatialized 3D audio, a head-mounted display, a data glove, and speech recognition and synthesis, creating one of the earliest multimodal VR systems in which trainees could interact using voice commands and receive audio feedback within a fully simulated environment (Fisher et al. 1989). During the 1990s, advances in computational power enabled more reliable automatic speech recognition (ASR) and more sophisticated VR rendering. Bohn and Krüger (1995) developed a speaker-independent, word-level recognition system based on neural network models designed to operate in immersive environments. Their approach employed a multilayer perceptron (MLP) trained on spectral feature vectors extracted from short-time windows of spoken input.

Other early approaches explored the integration of spoken dialogue with virtual agents. McGlashan (1995, 1996) examined how a spoken language interface could augment direct manipulation in VR and presented prototypes in which virtual characters could respond to user-issued speech commands, enabling basic conversational interaction within the environment. This work focused on combining speech input with graphical manipulation so that users could control objects or trigger actions either through physical interaction or verbal instruction. The system relied on a domain-specific grammar and semantic interpreter to map spoken utterances to agent actions, allowing users to execute commands such as object selection, movement, or the initiation of predefined behaviors.

Early pedagogical agents such as STEVE (SOAR Training Expert for Virtual Environments) integrated speech synthesis and recognition to teach procedural skills within dynamic 3D simulations. STEVE was developed as an intelligent tutoring agent capable of instructing trainees in complex tasks, including equipment operation, naval maintenance, and multi-step assembly. The system operated within a fully navigable 3D virtual environment, where the agent appeared as an embodied character capable of demonstrating actions, providing verbal explanations, and monitoring learner performance (Johnson, Rickel, and Lester 1998). Building on these principles, the Intelligent Virtual Environment for Language Learning (IVELL) (Johnson, Marsella, and Vilhjálmsson 2004) adopted embodied conversational agents that incorporated dialogue management, multimodal feedback, and real-time interaction to support second-language acquisition.

Figure 1.1: The Cave Automatic Virtual Environment (CAVE) at the Electronic Visualization Laboratory (EVL), University of Illinois at Chicago. Photograph by Dave Pape (self-timed capture), Public Domain, via Wikimedia Commons: https://commons.wikimedia.org/w/index.php?curid=868395.

Whereas STEVE relied on the SOAR cognitive architecture[2] to model instructional strategies, IVELL extended the concept of intelligent pedagogy by embedding conversational capabilities directly within the agent, enabling learners to practice linguistic structures, receive corrective feedback, and engage in contextually grounded dialogue within an immersive virtual scenario.

Healthcare applications further expanded these concepts by integrating speech-driven interaction into clinically oriented virtual environments. Greta is an embodied conversational agent designed to function as a virtual patient within immersive clinical training scenarios (Ochs and Pelachaud 2017). The system incorporates domain-adapted automatic speech recognition to process medically relevant vocabulary and trainee phrasing, enabling medical students to practice tasks such as delivering bad news or conducting sensitive interviews. Implemented using CAVE immersive technology[3], Greta provides synchronized verbal, facial, and gestural feedback, allowing trainees to experience emotionally complex interactions that approximate real clinical encounters.

Therapeutic VR systems have further employed natural speech dialogue to support intervention and rehabilitation. Systems for anxiety and phobia treatment have used spoken interaction to

2. SOAR is a general cognitive architecture for modeling human cognition, problem solving, and decision making through production rules, working memory structures, and a unified problem-space representation (Laird, Rosenbloom, and Newell 1987).

3. The CAVE (Cave Automatic Virtual Environment) is a room-sized projection-based virtual reality system introduced in Cruz-Neira, Sandin, and DeFanti (1993). It presents stereoscopic 3D scenes projected on multiple surrounding walls, enabling immersive, large-scale visualization with natural spatial orientation.

guide patients through exposure-based exercises and to assess real-time emotional state (Maples-Keller et al. 2017; Anderson et al. 2013). Speech-enabled VR tools for speech therapy provide opportunities for articulation practice, prosody training, and social communication rehearsal within controlled, customizable environments (Halpern et al. 2019; Bartoli, Corradi, and Garzotto 2013). Virtual counseling and mental-health support platforms similarly have adopted natural-language dialogue to simulate patient–therapist sessions, offering structured conversational practice and affective feedback in scenarios such as motivational interviewing or cognitive–behavioral therapy (Rizzo et al. 2014; Lucas et al. 2014).

In media art, speech has long functioned both as an interface and as expressive material, with automatic speech recognition and real-time vocal analysis enabling new forms of interaction in virtual and immersive environments. A seminal example is *The Hidden Worlds of Noise and Voice*, developed by Levin and Lieberman (2004) during an artist residency at the Ars Electronica Futurelab[4]. The project represents one of the earliest publicly exhibited installations to employ real-time vocal analysis as a primary interaction modality within a head-mounted augmented reality system. Conceived as an "augmented-reality speech-visualization system," the work used optical see-through data glasses to overlay animated three-dimensional graphics directly into the participant's field of view. As visitors spoke or sang, the system performed continuous acoustic analysis, extracting features such as pitch, amplitude envelope, and spectral brightness. These parameters were mapped onto volumetric visual forms that appeared to emanate from the speaker's mouth, effectively transforming vocal expression into shared spatial phenomena. Up to six users could interact within this environment simultaneously, observing others' vocal traces as dynamically evolving forms in augmented reality, while other spectators observed the results on external screens (see Figure 1.2). *Hidden Worlds* was installed alongside *RE:MARK* (2002) at the Ars Electronica Center, where both works formed part of a unified technical and conceptual research framework. While *Hidden Worlds* explored the immersive, spatial embodiment of vocal sound through augmented reality and head-tracked visualization, *RE:MARK* investigated the symbolic dimension of speech by applying phoneme recognition and projecting textual and abstract graphical marks onto a shared display surface. Importantly, the two installations shared core computational infrastructure, including real-time speech analysis systems, audio processing hardware, and tracking technologies, demonstrating a modular approach to speech visualization research.

In gaming and immersive entertainment, early research demonstrated that speech could complement gesture and controller input in VR, particularly for hands-free command execution and object manipulation (Sharma, Drury, and Ferrin 2015). Spoken interaction enables a natural coordination and role embodiment within immersive environments. A widely documented commercial deployment of speech interaction in VR is *Star Trek: Bridge Crew*, which integrates structured voice command recognition allowing players to issue spoken instructions to AI crew members during cooperative gameplay (Ubisoft 2017). Similarly, mod-supported implementations of *The*

4. The Ars Electronica Futurelab is a research and development laboratory affiliated with Ars Electronica—an international platform for media art, digital culture, and technological research founded in Linz, Austria, in 1979—that focuses on experimental media technologies, interactive installations, and immersive environments, often in collaboration with artists, scientists, and industry partners.

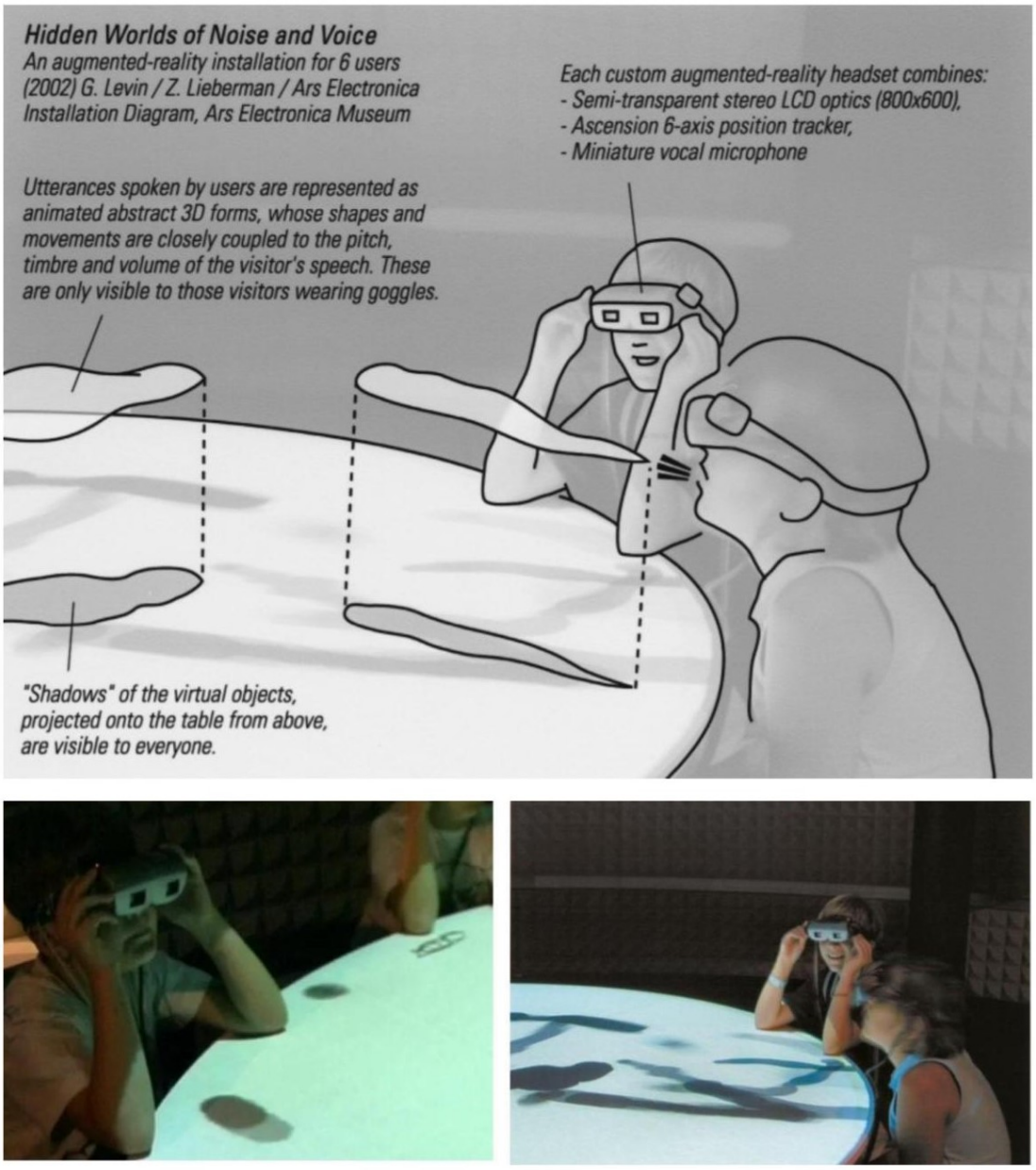
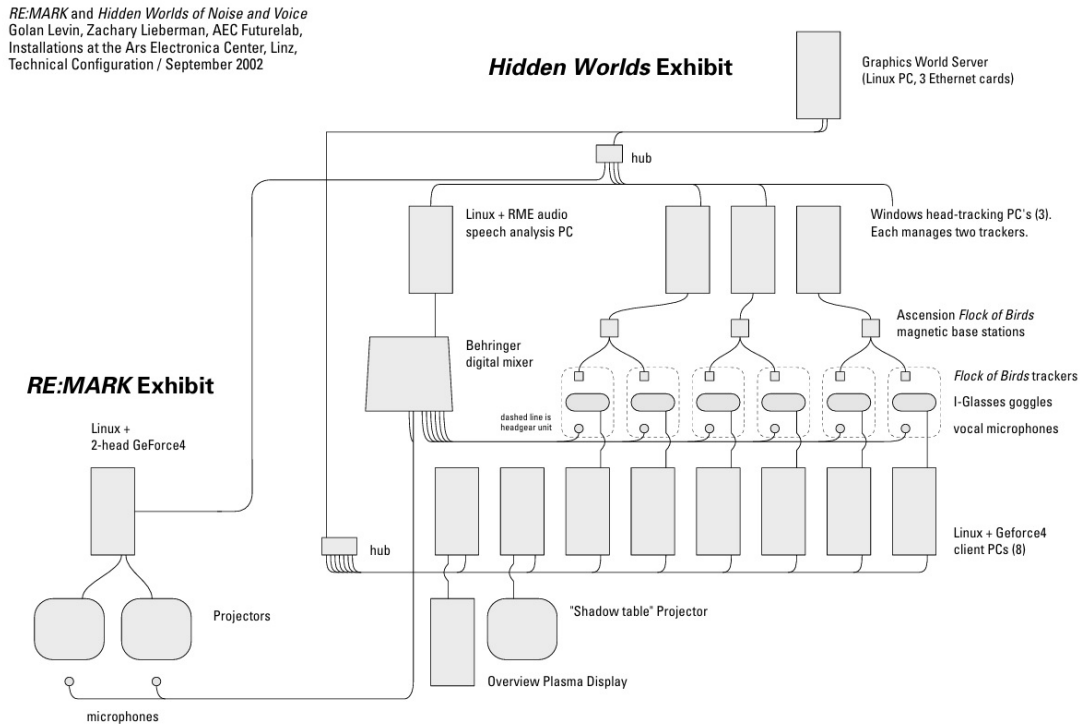


Figure 1.2: *Hidden Worlds of Noise and Voice* (2002), an augmented-reality installation by Golan Levin and Zachary Lieberman, developed at the Ars Electronica Futurelab and exhibited at the Ars Electronica Museum, Linz. The system employed custom optical see-through headsets with position tracking and embedded microphones to visualize users' vocal utterances as animated three-dimensional forms appearing to emanate from the speaker's mouth. While these volumetric graphics were visible only through the head-mounted displays, projected "shadows" of the virtual objects on the table surface allowed non-wearing spectators to observe the interaction. Source: https://www.flong.com/archive/projects/hwnv/index.html.



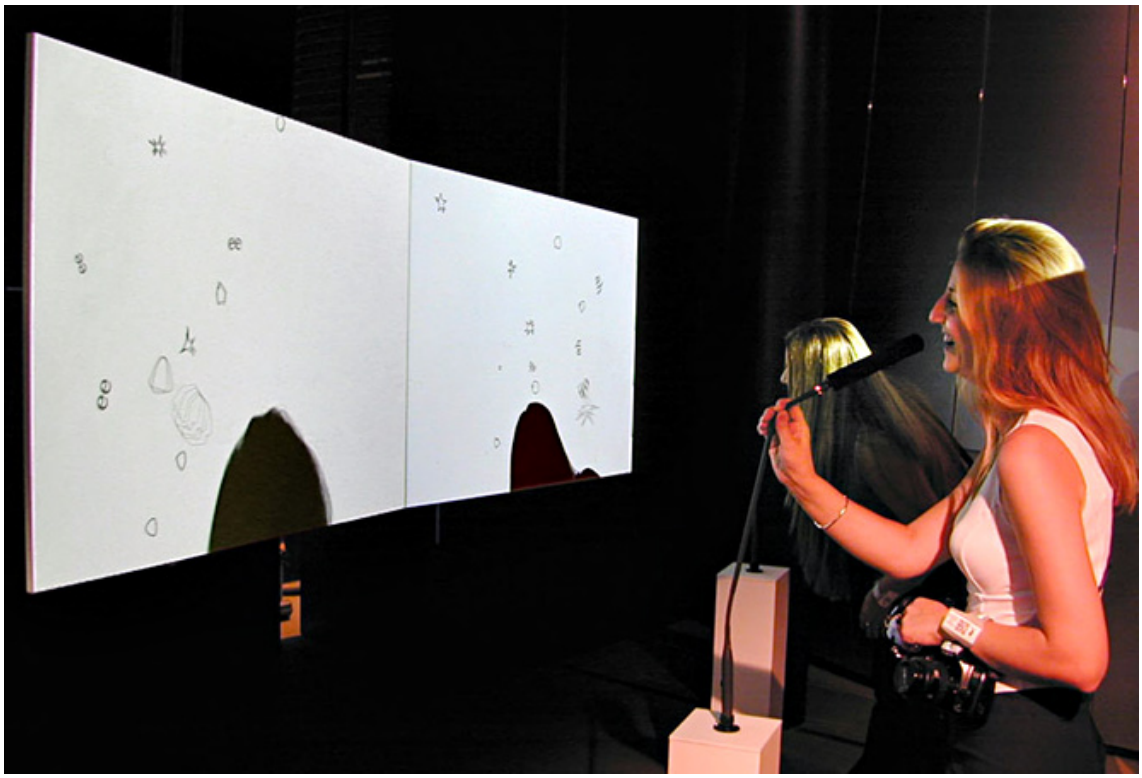


Figure 1.3: Technical configuration diagram of *RE:MARK* and *Hidden Worlds of Noise and Voice*, installed at the Ars Electronica Center (September 2002). The schematic (top) illustrates the shared computational infrastructure supporting both installations. The lower image shows *RE:MARK* in situ at the Ars Electronica Center. Source: Levin, G., Lieberman, Z., AEC Futurelab (2002). Available at https://personales.upv.es/moimacar/pages/catalog/03/image/golan/system.pdf.

Elder Scrolls V: Skyrim VR introduced speech-driven command input, allowing players to trigger in-game actions and interact with characters via spoken keywords (Bethesda Game Studios 2017). Conversational virtual characters combining automatic speech recognition, dialogue management, and embodied animation have been investigated as mechanisms for interactive storytelling (Mott, Lee, and Lester 2019; Lucas et al. 2014).

The *CALLAS* project (Vogt et al. 2009; Chryssanthou, Caridakis, Kabassi, et al. 2009) is an EU-funded framework that explored how multimodal emotion analysis; including vocal prosody, linguistic cues, and facial expressions, could be used to drive interactive storytelling and augmented-reality artworks. CALLAS aimed to create "emotion-rich media interfaces" and played a foundational role in establishing affective interaction strategies for artistic and cultural applications. A flagship demonstration of CALLAS was the *E-tree* installation, an augmented-reality sculpture whose behavior was shaped by the user's emotional voice input. The system combined real-time vocal emotion recognition with a generative 3D model rendered in AR. When participants spoke into a microphone, the system extracted prosodic features—such as pitch dynamics, energy contours, spectral tilt, and voicing probability—using the EmoVoice framework developed as part of CALLAS (Vogt and André 2008). These features were classified into discrete emotional categories (e.g., "calm," "angry," "happy," "sad") or mapped onto a continuous arousal–valence plane. The resulting affective estimates directly modulated the E-tree's visual appearance and behavior: branches elongated, shrank, or twisted; colors shifted across emotional gradients; and the density and motion of leaves reflected the user's expressive tone.

The *Auris* system (Sra et al. 2017) is an automated pipeline for generating immersive affective environments directly from musical input. Auris predicts a song's mood by extracting Mel-frequency cepstral coefficients (MFCCs) and feeding them into a gated recurrent unit (GRU)–based neural network classifier trained to assign one of four discrete mood categories—*happy*, *angry*, *sad*, or *calm*—corresponding to quadrants of the Russell–Thayer arousal–valence model. The system performs natural language processing on the song's lyrics, extracting noun phrases that serve as symbolic cues for environmental objects and scene elements. These affective (mood class) and semantic (lyrical phrase) representations jointly condition an image-generation module based on a multi-conditional PixelCNN architecture, which synthesizes base imagery reflecting both the emotional tone and textual content of the music. The generated images are subsequently enhanced using DeepDream convolutional processing,[5] producing highly textured visual materials (see Fig.1.5) that are mapped onto three-dimensional assets within a Unity-based VR environment. A genetic algorithm determines object placement and spatial composition, optimizing scene layout according to the detected mood category. A user evaluation study reported a strong correspondence between the intended musical mood and participants' perception of the generated VR scenes, indicating that the system successfully translated affective musical characteristics into spatial, visual, and atmospheric properties of immersive virtual

5. DeepDream is a gradient-ascent visualization technique introduced by Mordvintsev et al. that amplifies learned patterns within convolutional neural networks to produce hallucinatory and highly textured imagery (Mordvintsev, Olah, and Tyka 2015).

environments.

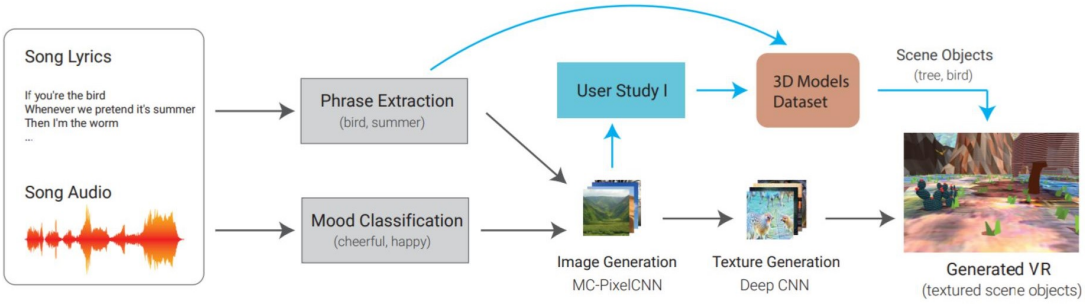


Figure 1.4: Overview of the *Auris* pipeline for generating affective virtual environments from music. Song audio is processed through MFCC-based mood classification, while song lyrics undergo noun-phrase extraction. The resulting affective (mood) and semantic (phrase) representations jointly condition a Mood-Conditional PixelCNN image generator. The synthesized images are transformed into textures and mapped onto 3D assets, producing a procedurally composed VR scene. Adapted from Sra et al. 2017.

Figure 1.5: Examples of VR environments generated by the *Auris* system. The scenes illustrate textured 3D objects whose materials are derived from DeepDream-enhanced images conditioned on musical mood and lyrical content. The resulting immersive environments visually encode affective characteristics inferred from the input songs. Adapted from Sra et al. 2017.

The SentimentVoice project (Ryu 2023) introduces a real-time affective computing pipeline that fuses acoustic emotion recognition with textual sentiment analysis to dynamically shape immersive VR theater environments. The system extracts prosodic cues to estimate the performer's arousal level, while transformer-based NLP models infer the sentiment polarity of spoken lines. These affective signals jointly drive generative stage elements including lighting, particle systems, and spatialized sound. Similarly, Khalaf et al. (2024) proposed a framework for emotion-aware content generation in the metaverse[6] that leverages natural language processing alongside a neuro-fuzzy emotion recognition engine. Their system recognizes users' emotional states and dynami-

6. The metaverse is generally described as a persistent, shared, and immersive virtual environment that integrates real-time interaction, embodied avatars, digital assets, and networked social presence across interconnected platforms.

cally adapts narrative elements and virtual scenes, illustrating the growing convergence of affective computing and generative techniques in immersive environments.

1.2 Problem Statement and Research Objectives

Affective computing research highlights a persistent imbalance in modality usage, with physiological and visual signals dominating emotion recognition studies, while speech and language are frequently treated as separate or sequential processing channels (García-Hernández et al. 2024). Although multimodal systems are increasingly reported, integrations between semantic sentiment analysis and acoustic–prosodic modeling remain methodologically fragmented and are rarely conceptualized as structurally independent yet interacting layers of affect representation. In particular, little attention has been given to computational frameworks capable of modeling systematic divergences between linguistic polarity and vocal affect—configurations that are central to figurative phenomena such as verbal irony.

Speech has long been recognized as a natural modality for interaction in virtual and immersive environments, complementing traditional interfaces such as controllers and graphical user interfaces (Bolt 1980; Oviatt 1999; Furness 1986). Within contemporary digital systems, speech functions not only as a command interface but also as a carrier of affect, intention, and expressive nuance. Beyond its instrumental role, speech may be conceptualized as an agential component within a broader assemblage (*agencement*) composed of bodies, technologies, affects, and signs (Deleuze and Guattari 1987, 1983). In this sense, vocal expression participates in the coproduction of meaning, shaping how computational systems respond, generate content, and structure experience.

Speech communicates across multiple interdependent representational layers of meaning (Mehrabian and Wiener 1967). Divergences between these layers give rise to emotional ambiguities, rhetorical tension, and attitudinal positioning. For instance, a sarcastic speaker may utter "Great job, I love this environment" while being notably disappointed or angry, expressing a negative prosodic attitude that contradicts the positive lexical sentiment. Conversely, in kind irony, a speaker might say "This is a terrible spot" while being happy, signaling a positive tone of voice despite the negative lexical content. In both cases, meaning emerges from the structured divergence between semantic polarity and prosodic valence.

However, contemporary speech interfaces typically rely on pipeline architectures in which the acoustic signal is first transcribed into discrete symbolic representations via automatic speech recognition. Downstream natural language processing components then operate primarily on textual tokens or embeddings derived from them. While this abstraction is effective for semantic interpretation and task execution, it reduces the high-dimensional acoustic signal to a constrained

It extends earlier virtual world paradigms by combining immersive VR/AR interfaces with social computing and decentralized infrastructures (Lee et al. 2021; Dionisio, Burns III, and Gilbert 2013).

symbolic sequence. In doing so, it frequently marginalizes—or entirely discards—prosodic variation that conveys subjectivity, emotional complexity, and the tension that arises when how something is said reshapes what is said.

Despite significant advances in speech emotion recognition and natural language processing, current systems remain poorly equipped to interpret expressive ambiguity. Conventional SER models typically map acoustic features to discrete emotional categories, whereas NLP-based sentiment analysis models infer affective polarity from lexical content alone. Neither modality, when used in isolation, is sufficient to capture ironic intent in speech. What remains lacking is a unified multimodal computational framework capable of representing and operationalizing the structured mismatch between prosodic and semantic cues that renders irony perceptible, and of proposing strategies for constructing virtual environments that deliberately diverge in audiovisual cues to generate ironic configurations.

From a computational perspective, progress is constrained by the scarcity of annotated datasets explicitly designed to capture ironic speech. Existing corpora (see Table 2.11) provide only limited coverage of systematically controlled contrasts between sincere, kind-ironic, and sarcastic speech acts. Moreover, few datasets offer consistent annotation protocols that support robust comparison across speakers, genders, and contextual conditions, thereby restricting both model interpretability and the reliability of quantitative evaluation.

From a digital media and creative systems perspective, this limitation exposes a deeper misalignment between contemporary computational approaches to affect and the ways in which expressive contradiction has been historically explored in artistic practice. In literary and performative traditions, irony does not function as a discrete emotional category to be classified, but rather as a relational and structural phenomenon emerging from tension, ambiguity, and contextual interplay between expressive layers.

From a critical perspective, the problem of computational irony can be situated at the intersection of philosophy, media theory, and artistic practice. *The Ironic Machines* operates simultaneously as an artwork and as a critical apparatus for investigating how computational systems might engage with forms of meaning that are inherently fragmented, contradictory, and unstable.

Central Research Problem

Given these limitations, the central research problem addressed in this thesis is formulated through two interdependent inquiries:

1. **The Analytical Problem:** How can prosodic irony be computationally inferred as a measurable divergence between semantic sentiment analysis and acoustic speech emotion recognition categorical predictions?

2. **The Generative Problem:** How can this cross-modal divergence be operationalized as a structured mapping strategy to drive the generation of adaptive affective virtual environments?

1.3 Research Questions

Guided by the methodological framework described in Chapter 3, this thesis addresses the following research questions:

1. **Can multimodal speech emotion recognition be used to detect and characterize prosodic verbal irony?** Specifically, can contradictions between semantic sentiment predictions and acoustic–prosodic emotion predictions serve as reliable indicators of kind irony and sarcasm?

2. **Which prosodic features most clearly differentiate ironic speech from sincere speech?** How do these features behave across ironic configurations and genders, and which dimensions remain stable or variable?

3. **How can bimodal emotional representations be mapped into affective virtual environments?** What strategies allow speech-derived emotional profiles to meaningfully drive audiovisual generation?

4. **Can audiovisual incongruence function as an analogue of verbal irony?** Does the deliberate opposition between musical and visual emotional profiles produce perceptible ironic structures in immersive environments?

Operational Tasks

To answer these questions, the research undertakes the following tasks:

- Design and implement a bimodal analysis pipeline integrating semantic sentiment analysis and acoustic speech emotion recognition.
- Formalize irony as a measurable divergence between semantic and prosodic emotional predictions.
- Extract, statistically analyze, and compare prosodic features across controlled ironic and sincere speech samples.
- Develop emotion-informed prompting strategies for music and image generation.
- Construct and perceptually evaluate affective virtual environments that express irony through structured audiovisual contrast.

Task	Addresses
Design a bimodal pipeline integrating sentiment analysis and acoustic SER	RQ1
Formalize prosodic irony as a measurable semantic–prosodic divergence	RQ1, RQ2
Extract and statistically compare prosodic features across sincere vs. ironic speech	RQ2
Develop emotion-informed prompting strategies for music and image generation	RQ3
Build and perceptually evaluate AVEs expressing irony through audiovisual contrast	RQ4

Table 1.1: Operational Tasks and Corresponding Research Questions

1.4 Research Objectives

To address this problem, the research pursues the following objectives:

1. **To model the divergence between semantic sentiment and acoustic prosody.** By combining transformer-based sentiment analysis with speech emotion recognition outputs, the research explicitly proposes prosodic irony as a measurable mismatch between lexical polarity and prosodic affect.

2. **To characterize prosodic features associated with ironic and sincere speech.** This involves identifying acoustic–prosodic patterns, such as pitch behavior, intensity, voice quality, and temporal structure, that differentiate kind irony and sarcasm from their sincere counterparts, using controlled subsets of annotated speech data.

3. **To translate ironic speech representations into affective virtual environments.** The research investigates how bimodal emotional profiles can be mapped onto generative audiovisual systems, enabling virtual environments that express irony through deliberate audiovisual opposition.

4. **To situate creative practice as a methodological instrument.** Through iterative development of artistic and technical artifacts, the research treats creative systems as experimental probes that test, reveal, and refine computational assumptions about affect, irony, and multimodal interaction.

1.5 Research Contributions

The project was disseminated through peer-reviewed research venues, public demos, and curated exhibitions across multimedia, music-AI, and immersive art contexts:

Main Publications

1. Forero, J., Bernardes, G., & Mendes, M. (2022). *Emotional Machines: Toward Affective Virtual Environments*. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM '22)*, pp. 7237–7238.
   DOI: https://doi.org/10.1145/3503161.3549973

2. Forero, J., Bernardes, G., & Mendes, M. (2023).[7] *Desiring Machines and Affective Virtual Environments*. In *ArtsIT, Interactivity and Game Creation* (ed. A. L. Brooks), pp. 405–414. Springer.
   DOI: https://doi.org/10.1007/978-3-031-28993-4_28

3. Forero, J., Mendes, M., & Bernardes, G. (2023). *En train d'oublier: toward affective virtual environments*. In *Proceedings of the 20th International Conference on Culture and Computer Science (KUI '23)*. ACM.
   DOI: https://doi.org/10.1145/3623462.3623469

4. Forero, J., Bernardes, G., & Mendes, M. (2023). *Are Words Enough? On the Semantic Conditioning of Affective Music Generation*. *arXiv preprint arXiv:2311.03624*.
   URL: https://arxiv.org/abs/2311.03624
   (Zenodo record / DOI): https://doi.org/10.5281/zenodo.8328584

5. Braga, F., Forero, J., & Bernardes, G. (2024). *Statistical Analysis of Musical Features for Emotional Semantic Differentiation in Human and AI Databases*. In *Proceedings of the 21st Sound and Music Computing Conference (SMC 2024)*.
   DOI: https://doi.org/10.5281/zenodo.14337968

6. Forero, J. Rodriguez, Bernardes, G., & Mendes, M. (2025). *The "What" Space: Prosodic Variability and Affective Virtual Environments*. In *Proceedings of the 16th International Conference on Computational Creativity (ICCC '25)*.
   URL (Proceedings): https://computationalcreativity.net/iccc25/accepted-papers/

7. Forero, J. Rodriguez, Bernardes, G., & Mendes, M. (2025). *The Ironic Machines*. Accepted / presented in the *12th International Conference on Digital and Interactive Arts (ARTECH 2025)*.
   DOI: https://doi.org/10.1145/3773699.3774674

Public presentations

- **ACM Multimedia 2022 — Interactive Art / Demo (Lisbon, Portugal).** *Emotional Machines: Toward Affective Virtual Environments* was presented within the ACM MM 2022

7. The ArtsIT conference edition is commonly referred to as ArtsIT 2022, but the proceedings volume is published in 2023.

context in Lisbon, articulating the system as an interactive installation that generates affective virtual environments through spoken language and speech emotion recognition.
DOI: https://doi.org/10.1145/3503161.3549973

- **Open Day CTM (INESC TEC) 2023 — Public Demo (Porto, Portugal).** The system was presented as a live demonstration under the CTM "Demos & Port of Honour" programme, listed as *Emotional Machines* (Jorge Forero, Multimedia Communications Technologies).
  Event page: https://opendayctm.inesctec.pt/

- **AIMC 2023 — Demo Sessions (University of Sussex, UK).** *Emotional Machines* was presented in the demo programme of the International Conference on Artificial Intelligence and Music Creativity (AIMC 2023), situating the system within discussions on AI-assisted music creativity and affective audiovisual instruments.
  Demo page: https://aimc2023.pubpub.org/pub/chx6olpn/release/1
  AIMC 2023 conference info (dates/location): https://aimusiccreativity.org/conferences/

- **XR Festival Florence 2024 — Immersive Exhibition (Florence, Italy).** *Terra Australis Ignota* was exhibited as part of XR Festival Florence (13–16 June 2024), hosted at the *Cattedrale dell'Immagine*. The festival programme frames the work within immersive audiovisual and XR storytelling.
  Festival page: https://www.caphe.space/xr-festival-florence-2024/

- **Noite Europeia dos Investigadores 2024 — Public Exhibition (Lisbon, Portugal).** A subsequent public-facing installation of *Terra Australis Ignota* was presented in Lisbon within the programme "Imersão: vídeo arte na investigação académica", explicitly listing the work and author.
  Programme: belasartes.ulisboa.pt/imersao-video-arte

- **Antología de Literatura Digital Latinoamericana (Lit(e)Lat) Vol. 2.0 — Digital Literature Context.** *Abandoned Memories* appears as a digital literature work in the Lit(e)Lat anthology, framing the project at the intersection of language, AI image processing, and virtual environments.
  Entry: https://antologia.litelat.net/volumen2/es/05_Abandoned_Memories.html

- **Loop Art Critique / Onland — Metaverse Exhibition Context (Online).** The practice also intersects with metaverse-based exhibition formats through Loop Art Critique / Onland, a programme combining critique residency structures with public exhibitions in a browser-based virtual space.
  Platform: https://loop.onland.io/
  Exhibition space: https://verse.loop.onland.io/q3JJ8bv/13th-ambiance-loop-art-critique

- **14th Algarve Design Meeting 2025 — Videomapping (Faro, Portugal).** *Bello* was presented as part of the interdisciplinary programme of the 14th Algarve Design Meeting,

an annual international design event hosted at Fábrica da Cerveja in Faro (19–24 May 2025). The programme included exhibitions, videomapping, outdoor installations, and design projects integrated within the urban context of Faro, situating the work at the intersection of audiovisual media, spatial design, and computational artistic practice.
Event archive: https://algarvedesignmeeting.ualg.pt/pt/adm/anteriores
14th ADM description: https://www.ualg.pt/14o-algarve-design-meeting

1.6 Thesis Structure

This dissertation is organized into eight chapters, each addressing a key dimension of the research on multimodal modelling of prosodic irony and its integration into interactive affective environments.

Chapter 2 — Background and State of the Art

Chapter 2 establishes the theoretical and technical scaffolding for the research by synthesizing literature across multiple disciplines. It begins with a critical review of emotional representation frameworks, contrasting categorical models (e.g., Ekman's discrete emotions) with dimensional models defined by valence and arousal. The chapter provides an extensive survey of state-of-the-art methods in Sentiment Analysis (SA) and Speech Emotion Recognition (SER), detailing standard datasets, machine learning architectures, and the historical evolution of performance metrics in these fields.

A central focus is placed on bimodal speech emotion recognition, specifically investigating the functional divergence between the sentiment analysis of the literal semantic content and the emotional acoustic prediction—a hallmark of figurative language such as kind irony and sarcasm. Furthermore, the chapter reviews current paradigms in affective algorithmic composition and generative systems, contextualizing the development of Affective Virtual Environments (AVE) and the visual semiotics required to translate computational affect into immersive digital experiences.

Chapter 3 — Methodology

Chapter 3 delineates the interdisciplinary methodological framework of the study, which operates at the intersection of media art and computer science. Grounded in practice-based research, the chapter details the iterative design and implementation of an end-to-end system that integrates automatic emotion recognition with the generation of Affective Virtual Environments (AVE).

The technical methodology is presented in two primary stages:

- **Bimodal Emotion Recognition:** The development of a dual-stream pipeline for Sentiment Analysis (SA) and Speech Emotion Recognition (SER). This includes the extraction of semantic polarity from text and acoustic-prosodic features (such as pitch and intensity) from speech signals.

- **Ironic Intent Assessment:** The implementation of a contrastive analysis framework that uses perceptual evaluation and statistical inference—specifically the Mann–Whitney U test and Cliff's delta—to identify significant prosodic differentiators between sincere and ironic speech.

The chapter further explores the creative methodology used to map these emotional predictions into virtual space. This involves a systematic profiling of musical and visual databases to establish high-, mid-, and low-level semantic prompting strategies for AI-guided generative systems. Finally, the methodology encompasses a series of nine creative iterations that function as both experimental case studies and autonomous artistic inquiries, documenting the transition from technical inference to immersive affective experience.

Chapter 4 — Results: Bimodal Speech Emotion Recognition

Chapter 4 presents the experimental results and performance evaluations of the bimodal Speech Emotion Recognition (SER) system, focusing on the individual and combined effectiveness of semantic and acoustic layers. The chapter first details the implementation of Sentiment Analysis (SA) models for textual content, contrasting traditional architectures like the Rocchio classifier with advanced deep learning approaches, specifically Long Short-Term Memory (LSTM) and Bidirectional Encoder Representations from Transformers (BERT). These models establish the baseline for literal sentiment detection.

The second half of the chapter focuses on the acoustic dimension, evaluating Multi-Layer Perceptron (MLP) and Convolutional Neural Network (CNN) architectures trained on the eGeMAPS and MFCC feature sets. Key findings highlight that eGeMAPS functional features significantly outperform MFCCs in emotional classification accuracy (71.18% vs. 59.38%). Furthermore, the chapter documents the critical role of waveform-level data augmentation—including time stretching and pitch shifting—in mitigating overfitting and enhancing the model's ability to generalize across diverse vocal expressions. The chapter concludes by synthesizing these unimodal results into a roadmap for detecting irony through cross-modal incongruity.

Chapter 5 — Results: Emotional Ambiguity and Verbal Irony

Chapter 5 investigates the emergence of irony through the functional divergence between semantic sentiment and acoustic emotional valence. Utilizing the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database, the chapter provides a rigorous statistical and exploratory analysis of how identical textual content can be perceived as different emotional states—such as frustration or neutral—depending on paralinguistic and situational markers.

The chapter details the identification of significant acoustic-prosodic features that differentiate "kind irony" and "sarcasm" from sincere expressions. Central to this analysis is a series of contrastive experiments, including the comparison of kind irony versus sincere anger and sarcasm versus sincere happiness. These technical findings are augmented by a human-centric perceptual evaluation conducted via the Podonos platform, which quantifies the degree of perceived irony and

correlates it with machine-learning predictions. Ultimately, the results presented here establish irony not as a categorical emotion, but as a configurational phenomenon grounded in measurable vocal behavior and cross-modal incongruity.

Chapter 6 — Affective Virtual Environment Generation

Chapter 6 presents a data-driven approach to the generation of affective virtual environments by translating empirically derived audiovisual features into structured generative prompts. The chapter investigates how low-level musical and visual descriptors—such as pitch salience, harmonic entropy, onset rate, color distributions, luminance, and textural complexity—contribute to the differentiation of emotional categories within the valence–arousal space. Using the 4Q Audio Emotion and Emotion6 datasets, the analysis identifies statistically significant features that reliably discriminate between emotional states, including onset density, dissonance, and spectral centroid in music, as well as brightness, saturation, entropy, and edge density in images.

Building on these findings, the chapter introduces a three-level prompt engineering framework designed to guide generative AI systems with increasing degrees of specificity. This framework progresses from high-level emotional labels, to mid-level semantic descriptors grounded in statistically significant features, and finally to low-level parametric prompts derived from empirical feature distributions (e.g., approximate BPM, spectral centroid values, or HSV brightness levels). This hierarchical structure enables a more controlled and interpretable mapping between affective intent and generated audiovisual output.

The framework is further applied to the construction of multimodal environments that combine generated music and images, including configurations that explore cross-modal emotional divergence, such as irony and sarcasm. A perceptual evaluation suggests that feature-informed prompts tend to improve audiovisual coherence compared to purely conceptual prompts, although the results remain exploratory due to sample size limitations. Overall, the chapter demonstrates how empirically grounded feature profiling can serve as an effective intermediate representation for designing affective virtual environments, while also highlighting the current constraints of prompt-based control in generative models.

Chapter 7 — Creative Practice

Chapter 7 documents the realization of the research through a series of nine technological and artistic iterations, which function both as experimental validations of the proposed system and as autonomous creative inquiries. These works explore the boundary between human affect and machine-mediated representation, operating within a recurring narrative of disorientation and the speculative construct of “home” as an imaginary locus. The chapter traces the progressive development of a set of interconnected interactive systems. Early iterations establish the foundational integration of speech, sentiment analysis, and real-time audiovisual generation, while later iterations expand this framework toward multimodal affective environments and the exploration of

ambiguity, prosody, and irony. The chapter details the evolution of various types of artifacts, including:

- **Early Prototypes:** Browser-based and Unity systems that integrate real-time speech-to-text transcription with sentiment lexicons and rule-based commands to dynamically modulate sound synthesis and manipulate virtual objects.
- **Bimodal Generative Systems:** Iterations developed under the *Emotional Machines* framework, combining semantic sentiment analysis, acoustic speech emotion recognition, generative image pipelines, and algorithmic music systems to construct immersive affective virtual environments driven by spoken input.
- **Geo-poetic and Speculative Environments:** Projects such as *Abandoned Memories*, *Veintitantos Vientos*, *En train d'oublier*, and *Terra Australis Ignota*, which explore the relationship between place, memory, and affect through generative cartographies, diffusion-based imagery, and speech-driven transformations of real and imagined landscapes.
- **Exploratory Affective Systems:** Projects such as *The What Space*, which investigate emotional ambiguity arising from prosodic variation through data-driven visualizations, network structures, and real-time emotion-to-texture translation.
- **Ironic Audiovisual Systems:** Final iterations such as *The Ironic Machine[s]* and *Bello*, which operationalize irony as a structured divergence between modalities, generating affective virtual environments through controlled audiovisual incongruence.

Chapter 8 — Conclusion and Future Work

Chapter 8 serves as the final synthesis of the research, evaluating the extent to which the experimental results and creative outputs validate the central hypothesis. It reflects on the technical efficacy of the bimodal Speech Emotion Recognition (SER) pipeline and the aesthetic viability of the generative Affective Virtual Environments (AVE) in conveying the nuances of prosodic irony. By bridging the quantitative findings of Chapters 4 and 5 with the practice-based reflections of Chapter 7, the conclusion articulates a new framework for understanding irony as a measurable cross-modal divergence in digital media.

The dissertation concludes by identifying the limitations encountered during the research—specifically regarding the scarcity of naturalistic, non-acted ironic datasets—and proposes trajectories for future work. These include the potential for real-time, low-latency ironic intent detection in social robotics and the expansion of the generative pipeline to include haptic and spatial audio affordances, further pushing the boundaries of immersive affective communication.

Chapter 2

Background and State of the Art

This chapter provides the theoretical, computational, and artistic foundations that underpin the development of the *Ironic Machines*. As the most extensive chapter of this dissertation, it establishes the interdisciplinary framework required to design a system capable of generating immersive virtual environments from both semantic content and acoustic–prosodic expression. The central objective of the proposed AVE system is to map bimodal speech information into adaptive affective environments. Within this framework, semantic polarity extracted from language and emotional valence inferred from vocal prosody jointly inform the generation of audiovisual structures. These environments are constructed through diverse computational strategies, including affective music generation, algorithmic composition, and generative visual modeling. The resulting immersive spaces dynamically adapt to the affective configurations inferred from speech input, including potential divergences between modalities.

To support this objective, section 2.2 and section 2.3 review the state of the art in emotion recognition from text (Sentiment Analysis) and from speech (Speech Emotion Recognition), respectively. These sections establish the computational mechanisms required to model semantic and acoustic affective signals as measurable and comparable dimensions. The theoretical foundations of emotion measurement that underlie these approaches are introduced in section 2.1, which surveys categorical and dimensional models of affect. Building on this framework, section 2.4 introduces bimodal speech emotion recognition and positions irony as a phenomenon of emotional divergence. In this work, irony is conceptualized as a measurable discrepancy between semantic sentiment and prosodic emotional valence. The section reviews linguistic, pragmatic, and acoustic research on ironic and sarcastic speech, identifying prosodic markers and multimodal patterns that inform the operational definition adopted here.

Finally, section 2.5 and section 2.6 present the background on affective music generation and affective virtual environment synthesis. These sections survey rule-based, evolutionary, neural, and multimodal generative approaches for producing emotionally modulated sound and visual environments. Together, they outline the computational and aesthetic strategies through which affective states and, critically, cross-modal divergences, can be rendered as immersive audiovisual experiences.

2.1 Measuring Emotions

The James–Lange theory proposed that emotions arise from the perception of physiological and motor responses to biologically significant stimuli (James 1884; Lange 1885). Cannon (1927) later challenged this view, arguing that emotional experience often precedes behavior, that visceral changes are too slow and undifferentiated to account for specific emotions, and that bodily responses are neither sufficient nor necessary for emotional experience. Building on this critique, the Cannon–Bard theory posited that emotions originate in the central nervous system, with subjective feeling and physiological arousal occurring simultaneously rather than sequentially (Cannon and Bard 1929). Extending this shift toward activation-based accounts, Duffy (1957) introduced the concept of arousal as a continuum of physiological activation ranging from deep sleep to extreme excitement.

Schachter and Singer's Two-Factor Theory proposed that emotional experience arises from the combination of undifferentiated physiological arousal (intensity) and the cognitive appraisal of situational cues (quality or label) (Schachter and Singer 1962). Mandler's interruption theory similarly argued that autonomic arousal is a nonspecific response to the interruption of ongoing cognitive activity; this arousal is necessary but not sufficient for emotion, which emerges when the individual interprets the cause and significance of the interruption (Mandler 1975, 1984). Appraisal theories, introduced by Arnold (1960) and further developed by Frijda (1986), Lazarus (1991), and Scherer (1999), emphasize emotion as a complex response to significant events, grounded in the appraisal of an intentional object (person, event, or situation). This appraisal process attributes meaning and affective value, thereby initiating an emotional state that includes both an evaluative component and an action tendency.

Juslin and Sloboda propose that emotions constitute one component within the broader domain of affect, which also includes moods, preferences, and enduring personality dispositions (Juslin and Sloboda 2010). A common distinction concerns duration and intensity: emotions tend to be brief and relatively intense (Parkinson 1995), whereas moods are generally more diffuse and sustained over longer periods of time (Eysenck and Keane 2000). Scholars have also emphasized the importance of distinguishing between *felt*, *expressed*, and *perceived* emotions. Russell (2003) argues that perceived emotions—those attributed by an observer to external stimuli—do not necessarily correspond to the emotions actually felt by an individual, since perception relies on inferences drawn from expressive cues, whereas felt emotion emerges from internal appraisal and core affective processes. Scherer (2004, 2005) further clarifies this distinction by separating expressed emotion (the signal emitted by a vocal, facial, or musical display) from perceived emotion (the meaning decoded by a listener or observer), noting that neither is guaranteed to align with the subjective emotional experience of the sender. Barrett, Satpute, and Bliss-Moreau (2019) similarly argues that categorical interpretations of perceived emotion often differ from the experiential organization of felt emotion, since emotion perception depends on conceptual knowledge and contextual inference. Complementing these perspectives, neuroscientific evidence reviewed by Saarimäki (2021) demonstrates that perceived and felt emotions are represented by partially

distinct cerebral topographies.

Two paradigmatic frames of reference dominate emotion research: categorical and dimensional models, although several alternatives have been proposed. Categorical models posit that emotional experiences can be represented as discrete entities, such as happiness, sadness, anger, fear, surprise, and disgust, each reflecting an evolutionarily grounded response pattern (Ekman 1992; Izard 1977). In contrast, dimensional models conceptualize emotions as positions within a continuous affective space, most commonly defined in terms of the orthogonal dimensions of valence and arousal (Russell 1980). Prototype theories further suggest that language and conceptual knowledge shape the formation of emotion categories: rather than being sharply bounded, emotions are organized around prototypical exemplars, with category membership determined by graded similarity (Shaver et al. 1987).

2.1.1 Categorical Frames

The idea that certain fundamental emotions provide the basis for more complex emotional states is consistent with Darwin's evolutionary view of emotion and has been further developed by various authors, as summarized in Table 2.1. Darwin proposed that many expressive behaviors originate as "serviceable associated habits," that is, bodily actions that were once adaptive and have persisted through evolution. Emotional expressions therefore serve functional purposes, such as protecting the body or preparing it for action, rather than merely communicating mood states. Darwin also argued that similarities in emotional expressions across humans and non-human animals provide evidence of a common evolutionary origin. Moreover, through his "principle of antithesis," he suggested that opposite emotions tend to produce opposite forms of expression, as seen in the contrasting postures associated with joy and sadness (Darwin -Ekman edition 1998). [1]

Following these evolutionary proposals, Ekman developed his influential classification of basic emotions, primarily based on characteristic facial expressions associated with each emotion, namely joy, sadness, anger, surprise, disgust, and fear (Ekman 1992, 2005). Cross-cultural studies conducted with non-Western populations demonstrated strong agreement in emotion recognition across cultures, although fear was occasionally confused with surprise due to overlapping facial features (Ekman and Friesen 1971). Ekman argued that these basic emotional expressions correspond to biologically innate emotions that serve adaptive functions related to individual and species survival and give rise to more complex emotional states. He later proposed contempt as a seventh basic emotion (Ekman 1992, 2005). Extending this evolutionary perspective, Izard proposed a theory of ten basic emotions, each characterized by an innate adaptive function, a distinct neuromuscular expressive pattern, and a specific subjective experience. These emotions include interest–excitement, joy, surprise, distress–anguish, anger, disgust, contempt, fear, shame, and guilt (Izard 1977, 1991).

1. Darwin's original work *The Expression of the Emotions in Man and Animals* was first published in 1872. The edition cited here corresponds to the modern annotated version edited by Paul Ekman, which includes historical context, critical commentary, and comparative analysis of Darwin's observations in light of contemporary affective science.

Panksepp (1982) proposed the hypothesis that a limited number of biologically grounded emotional action systems are rooted in subcortical brain circuits. In his early formulation, these systems were identified as Expectancy, Rage, Fear, and Panic. Subsequently, Panksepp (1998) expanded this framework and articulated a set of seven primary-process emotional categories, classified as SEEKING (or Expectancy), RAGE (anger), FEAR (anxiety), LUST (sexual excitement), CARE (nurturing), PANIC (separation distress or sadness), and PLAY (social joy).

Nesse (1990) argued that emotions can be understood as specialized states shaped by natural selection to cope with situations involving threats or opportunities. In this view, negative emotions are often recruited in contexts of threat and uncertainty, whereas positive emotions are more likely to arise in situations appraised as safe or advantageous. For example, anger can mobilize energy and motivation for assertive action aimed at resolving perceived obstacles or uncertainty, even when success is not guaranteed. Similarly, contentment is commonly associated with situations appraised as safe and satisfactory (Fredrickson 1998).

Theoretical Framework	**Discrete Emotion Taxonomy**
Ekman and Friesen 1971	Anger, happiness, fear, surprise, disgust, sadness
Plutchik 1980	Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise
Frijda, Manstead, and Oatley 1989	Desire, happiness, interest, surprise, wonder, sorrow
Ekman 1992	Happiness, surprise, fear, sadness, anger, disgust
Parrott 2001	Anger, fear, joy, love, sadness, surprise
Izard 2007	Joy, interest, fear, sadness, disgust, anger, guilt, shame, contempt, love, attachment
Grandjean, Sander, and Scherer 2008	Fear, pain, lust, pleasure, anger, joy, sadness, disgust
Jack, Garrod, and Schyns 2014	Fear, anger, joy, sadness
Gu et al. 2015	Fear, anger, joy, sadness
Junior and Wang 2016	Fear, anger, joy, sadness
Zheng et al. 2016	Fear, anger, joy, sadness
Gu et al. 2019	Happiness, sadness, fear, anger

Table 2.1: Taxonomies of primary emotion sets: a summary of foundational and contemporary classification models in affective science.

In later reflections on his work, Ekman noted that the number of emotion-related adjectives exceeds the number of basic emotions, suggesting that many lexical emotion terms may correspond to variants, blends, or different intensities within broader families of basic emotions (Ekman 1992, 2005).

2.1.1.1 The Affect Adjective Check List

Adjective-based instruments operationalize affective experience through lexical descriptors, reflecting the assumption that emotional states can be meaningfully accessed and differentiated via language. Common examples include the Multiple Affect Adjective Check List (MAACL) (Zuckerman et al. 1964), the Profile of Mood States (POMS) (McNair, Lorr, and Droppleman 1971),

and PANAS/PANAS-X (Watson, Clark, and Tellegen 1988; Watson and Clark 1994). These tools are especially useful when experimental tasks benefit from explicit semantic articulation of affect.

Although self-report measures are inherently influenced by subjective experience, introspective ability, and response biases (Allen and Gold 1986), they remain widely used in both laboratory and applied contexts due to their efficiency and interpretability. In particular, affective adjective norms such as the Affective Norms for English Words (ANEW) provide standardized ratings of emotional valence and arousal for large vocabularies of emotion-related terms, enabling the systematic use of language as a proxy for emotional meaning in experimental and computational settings. In their seminal paper, Watson and Tellegen (1985) conducted a critical review of the literature on the use of adjective checklists for measuring mood, investigating the extent to which different affective states can be adequately described using a range of lexical descriptors. Their analysis identified two primary dimensions of mood exhibiting bipolar characteristics.

2.1.2 Dimensional Frames

In contrast to categorical frameworks, in which emotional states are represented through the presence or absence of discrete emotion labels or adjectives, dimensional approaches conceptualize emotion as a continuous space in which affective states vary along one or more continuous dimensions. Rather than assigning an emotional experience to a single category, dimensional models describe emotional states as positions within a continuous affective continuum.

One of the earliest and most influential dimensional measurement techniques is the Semantic Differential (SD), originally introduced by Osgood (1953) and later formalized by Osgood, Suci, and Tannenbaum (1957) as a method for investigating the cross-cultural universality of meaning. Kerlinger (1973) defined the SD as "a method of observing and measuring the psychological meaning of concepts." The SD is a scaling instrument that enables the quantitative representation of judgments along continuous bipolar dimensions using polarized verbal anchors. Empirical studies employing the semantic differential have consistently shown that affective judgments tend to converge along three principal dimensions: Evaluation, Potency, and Activity. These dimensions parallel similar dimensional reductions proposed in emotion research, such as valence, arousal, and dominance (or potency) (Kleinen 1968; Nielzén and Cesarec 1981; Wedin 1972).

2.1.2.1 Rating scales

Dimensional rating scales are widely used in self-report methodologies to study perceived and induced emotional states. Prominent examples include the Differential Emotion Scale (DES) (Izard 1977) and the Positive and Negative Affect Schedule (PANAS), introduced by Watson, Clark, and Tellegen (1988), along with its extended version PANAS-X (Watson and Clark 1994). In the DES, participants rate sets of emotion-related words corresponding to basic emotions, reflecting a categorical organization embedded within a dimensional response format. In contrast, PANAS-X requires participants to evaluate affective concepts using graded intensity scales (e.g., "very slightly," "a little," "moderately," "quite a bit," "extremely"), thereby emphasizing continuous variation in

affective experience. The Profile of Mood States (POMS) similarly relies on adjective-based ratings using fixed Likert scales to assess transient mood dimensions such as tension, depression, anger, vigor, fatigue, and confusion (McNair, Lorr, and Droppleman 1971).

Another commonly used dimensional instrument is the Visual Analogue Scale (VAS), which captures affective intensity without imposing discrete categorical steps (Huskisson 1974). Typically implemented as a horizontal line of fixed length anchored by verbal descriptors at each end (e.g., "not at all joyful" to "extremely joyful"), the VAS allows participants to indicate the perceived intensity of their emotional state with high temporal and numerical resolution (Price et al. 1983; McCormack, Horne, and Sheather 1988).

2.1.2.2 Multidimensional Systems

The idea that emotional experience can be described within a multidimensional space was first articulated by Wundt (1905), who combined introspective, experimental, and physiological methods to study emotion. Wundt proposed that emotional states could be characterized along three independent dimensions: pleasantness–unpleasantness, rest–activation, and relaxation–attention. This foundational proposal influenced later multidimensional models, as shown in Table 2.2, including those developed by Plutchik (1960, 1982) and Russell (1980, 1983).

Russell's circumplex model of affect, introduced in 1980, represents one of the most influential dimensional models in contemporary emotion research. Based on similarity judgments of emotion terms and multidimensional scaling analysis, Russell (1980) identified two primary bipolar dimensions: valence (pleasantness–unpleasantness) and arousal (high–low activation). Within this framework, emotional states are represented as points in a two-dimensional space, with valence and arousal treated as orthogonal dimensions.

Several extensions and refinements of two-dimensional affect models have since been proposed. Thayer (1989), for example, subdivided the arousal dimension into energetic arousal and tense arousal, corresponding to distinct psychobiological systems associated with action readiness and stress-related activation, respectively.

Theoretical Source	**Dimensional Axes / Affective Factors**
Wundt 1905	Pleasant–unpleasant, excitation–calm, tension–relaxation
Schlosberg 1954	Pleasant–unpleasant, tension–relaxation, excitation–calm
Osgood, Suci, and Tannenbaum 1957	Evaluation (valence), potency, activity (arousal)
Russell 1980	Valence, arousal
Posner, Russell, and Peterson 2005	Valence, arousal
Grandjean, Sander, and Scherer 2008	Valence, arousal

Table 2.2: Evolution of multidimensional emotion models: A comparative overview of foundational axes used to map affective states.

In experimental studies employing dimensional models, participants are typically asked to rate emotional stimuli along valence and arousal dimensions using bipolar scales. Notably, even within dimensional paradigms, categorical emotion concepts are often implicitly present, either through

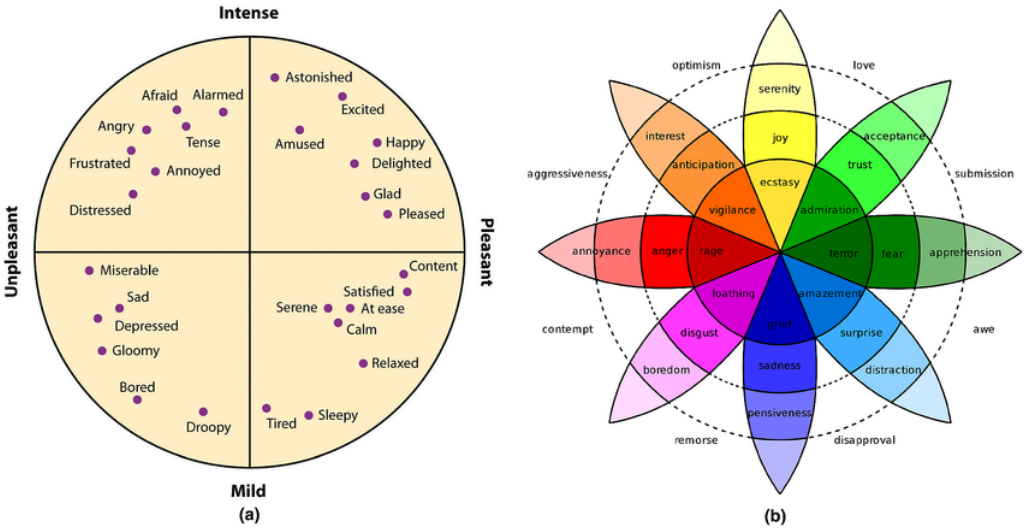


Figure 2.1: (a) Russell's Circumplex Model, illustrating the distribution of affective states along the dimensions of valence (pleasure) and arousal (activation). (b) Plutchik's Wheel of Emotions, depicting the intensity, relationships, and polarities of primary and secondary emotional categories.

the selection of emotion-eliciting stimuli or through the verbal anchors used at the scale extremes (e.g., labeling positive valence as "happy") (Baumgartner, Esslen, and Jäncke 2006; Gosselin, Kirouac, and Doré 2005; Peretz, Gagnon, and Bouchard 1997).

Dimensional models have been extensively applied in physiological and neuro-psychological studies of emotion, with particular emphasis on the valence dimension (Lang 1993). Davidson (1992, 1993) proposed a neuro-biological framework linking approach–avoidance mechanisms to positive and negative valence, suggesting partially lateralized neural substrates for these affective processes. Borod (1992, 1993) provided a comprehensive review of dimensional approaches in the neuro-psychology of emotion, further consolidating the relevance of dimensional representations for both experimental and applied research.

2.1.2.3 SAM and Other Pictorial/Grid Interfaces (Valence–Arousal–Dominance)

Within affective computing, an important family of instruments replaces verbal reports with pictorial or spatial interfaces. The Self-Assessment Manikin (SAM) is a non-verbal pictorial assessment technique that directly measures pleasure (valence) and arousal, and is commonly extended to include dominance as a third dimension (Bradley and Lang 1994). SAM scales efficiently across repeated trials and reduces linguistic load, making it particularly compatible with multimodal datasets and interactive experimental designs. Figure 2.2 illustrates the standard SAM interface.

A complementary rapid assessment tool is the Affect Grid, a single-item two-dimensional grid enabling quick placement of affect along pleasure (valence) and activation (arousal) (Russell,

1. Image source: https://www.researchgate.net/publication/236078116_Affective_auditory_stimuli_Adaptation_of_the_International_Affective_Digitized_Sounds_IADS-2_for_European_Portuguese/figures?lo=1. The original SAM instrument was introduced by Bradley and Lang (1994).

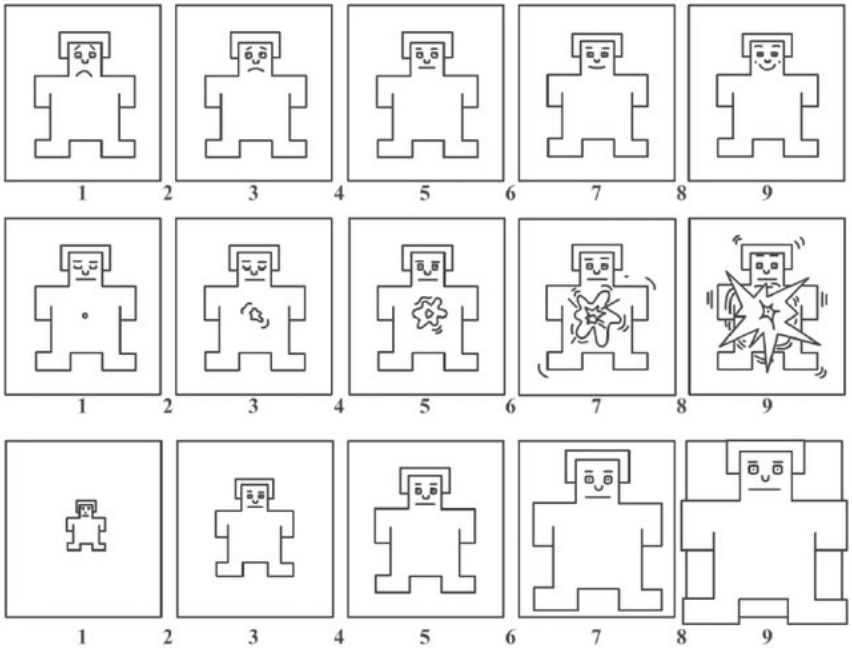


Figure 2.2: The Self-Assessment Manikin (SAM) interface (Bradley and Lang 1994) for measuring affect along the dimensions of valence (pleasure), arousal (activation), and dominance (control). The pictorial format enables rapid, language-independent emotional self-report. Image adapted from Sainz-de-Baranda Andujar et al. (2022).

Weiss, and Mendelsohn 1989). Such interfaces are particularly useful when emotional states must be sampled frequently without substantially interrupting an ongoing task.

2.1.3 Reliability, agreement, and quality control

Because emotion measurement is often mediated by annotators, observers, and instruments, it is crucial to report reliability and control for systematic biases. For categorical labels, chance-corrected agreement measures such as Cohen's kappa (Cohen 1960) and Fleiss' kappa for multiple raters (Fleiss 1971) are commonly used. For continuous or ordinal judgments, intraclass correlation coefficients (ICC) provide a principled framework for assessing rater reliability under different measurement assumptions (Shrout and Fleiss 1979). More generally, Krippendorff's approach foregrounds the conditions under which reliability coefficients should be interpreted as valid indices of data quality and is particularly appropriate when dealing with heterogeneous annotation designs (Krippendorff 2004).

2.2 Sentiment Analysis

Sentiment Analysis (SA) is a subfield of natural language processing concerned with identifying and characterizing subjective information in textual data, encompassing opinions, emotions, attitudes, and affective states. At its most basic, SA involves assigning sentiment polarity to text, distinguishing positive, negative, or neutral expressions, though modern approaches address considerably finer distinctions, including emotion categories, aspect-level sentiment, and implicit or figurative affect (Liu 2015; Pang and Lee 2008). Importantly, SA operates across multiple levels of granularity, from entire documents down to individual entities or aspects within a sentence, and its underlying goal is to model human affective expression in a form that is computationally tractable and practically useful.

Sentiment analysis gained prominence in the early 2000s, emerging in parallel with the rapid expansion of web-based platforms and the proliferation of large-scale user-generated content. Online reviews, forums, and social media produced vast corpora of opinionated and emotionally laden text, providing an unprecedented empirical foundation for computational approaches to affect (Turney 2002). Since then, SA has become an actively studied area in NLP, with applications spanning market research, customer relationship management, political communication, brand monitoring, and public health surveillance (Cambria et al. 2013). The field has undergone several major methodological transitions, from lexicon-based heuristics to supervised machine learning, and subsequently to deep neural networks and large pre-trained language models, each substantially expanding the range of linguistic phenomena that can be reliably modeled.

Beyond commercial and scientific applications, sentiment analysis has also been applied in artistic and research-creation contexts, where affective predictions are used as generative material rather than informational output. In the context of media art, sentiment analysis has proven valuable for interactive and immersive installations that leverage audience affect to dynamically shape artistic experience. One example is *T(w)peeper*, an interactive generative artwork that retrieves live tweets based on a user-selected keyword and performs real-time sentiment polarity classification. The resulting positive and negative affective values are mapped to contrasting visual parameters, gradually composing an evolving abstract image that reflects the collective mood of online discourse (Huang 2013). Another case is *HeartBeat*, an interactive public installation developed during the COVID-19 pandemic to visualize the collective emotional landscape of Canada in near-real time. The project, presented as a research-creation work by Akhtar (2021), collected large-scale Twitter data and applied sentiment and emotion analysis to surface prevailing affective patterns, including fear, hope, and grief, across time. These trends were rendered as a pulsing dynamic display, evoking a societal heartbeat whose rhythm was driven entirely by the emotional texture of public language. *HeartBeat* foregrounds the interpretive and political dimensions of affect quantification: choices about which emotions to extract, how to weight them, and how to render them visually are simultaneously technical and curatorial. These design choices reflect the interpretive dimension inherent in any affective quantification system, where technical decisions carry implicit assumptions about emotional categorization.

2.2.1 Sentiment Analysis Databases

Early sentiment analysis research focused primarily on datasets designed for binary polarity detection, categorizing texts as either positive or negative. The pioneering work of Pang, Lee, and Vaithyanathan (2002) introduced a corpus of movie reviews manually annotated for sentiment polarity, establishing the first widely adopted benchmark for sentiment classification algorithms. Subsequently, the IMDb reviews dataset (Maas et al. 2011) emerged as a larger and more representative resource, becoming a standard evaluation corpus owing to its broad coverage of informal and subjective linguistic expression in user-generated text.

Following early binary classification datasets, sentiment analysis research expanded toward more nuanced and context-sensitive tasks. The Stanford Sentiment Treebank (SST) advanced the field by introducing phrase-level annotation across five sentiment classes, ranging from very negative to very positive, encompassing approximately 11,800 sentences and 215,000 unique phrases when all parse tree nodes are included, significantly enhancing the ability to model linguistic subtlety and sentiment intensity (Socher et al. 2013). In parallel, the SemEval shared task series substantially enriched the field by providing diverse benchmarks for increasingly complex problems, including aspect-based sentiment analysis, in which sentiment is annotated with respect to specific entities or attributes within a text (Pontiki et al. 2016). SemEval-2024 Task 10 (EDiReF) introduced datasets for emotion discovery and emotion-flip reasoning within conversational and code-mixed dialogues, reflecting the growing research interest in dynamic, context-sensitive affective modeling (Kumar, Garg, and Singh 2024).

Sentiment analysis datasets have also expanded beyond purely textual data into multimodal settings, integrating acoustic and visual information alongside language. The CMU-MOSI (Zadeh et al. 2016) and CMU-MOSEI (Zadeh et al. 2018) corpora established the primary English-language benchmarks for continuous multimodal sentiment regression, while MELD (Poria et al. 2019) extended multimodal modeling to the emotion recognition in conversation setting. A significant addition to this landscape is CH-SIMS (Yu et al. 2020), a Chinese multimodal dataset that provides both unified multimodal and independent per-modality sentiment annotations, an important feature for studying cases in which expressed sentiment diverges across modalities. Together, these resources have driven the development of fusion architectures capable of capturing affective cues that no single modality can fully convey. In parallel, multilingual resources such as the Multilingual Amazon Reviews corpus (Keung et al. 2020) and the Cross-Lingual Natural Language Inference benchmark (XNLI) (Conneau et al. 2018) have broadened sentiment research into cross-lingual and language-agnostic settings.

More recently, sentiment analysis datasets have evolved to address complex emotional nuances and specialized linguistic contexts. GoEmotions (Demszky et al. 2020), a corpus of 58k Reddit comments annotated for 28 emotion categories (27 fine-grained emotions and neutral), represents a substantial departure from coarse polarity labeling toward psychologically grounded emotion taxonomies, and has become a widely adopted resource for fine-grained emotion classification research. The DynaSent dataset complements this by emphasizing contextualized annotation and an

explicit neutral category, directly targeting the limitations of binary polarity schemes in naturally occurring conversational text (Potts et al. 2021). Figurative language has also received dedicated attention, with the Sarcasm on Reddit corpus (Khodak, Saunshi, and Vodrahalli 2018) and the SemEval-2018 irony detection dataset (Van Hee, Lefever, and Hoste 2018) providing resources for modeling non-literal affect. In the domain of dialogue-based sentiment, Carvalho, Oliveira, and Silva (2022) introduced TwitterDialogueSAPT, a Portuguese corpus of customer service conversations with manual sentiment annotations, enabling the evaluation of models on domain-specific, linguistically varied interaction data. Collectively, these developments reflect a broader shift in the field toward datasets that capture affect as it actually occurs in context, ambiguous, figurative, and deeply shaped by conversational dynamics.

Dataset	**Year**	**Type**	**Classes**	**Modality**
Pang et al.	2002	Binary	2	Text
IMDb	2011	Binary	2	Text
SST	2013	Fine-grained	5	Text
SemEval (ABSA)	2016	Aspect-based	Multi	Text
Twitter SemEval	2017	Multi-class	3	Text
CMU-MOSI	2016	Continuous	Regression	Multimodal
CMU-MOSEI	2018	Continuous	Regression	Multimodal
MELD	2019	Multi-class	7	Multimodal
GoEmotions	2020	Fine-grained	28 (incl. neutral)	Text
CH-SIMS	2020	Continuous	Regression	Multimodal
DynaSent	2021	Ternary	3	Text
SemEval EDiReF	2024	Multi-class	Multi	Text

Table 2.3: Summary representative sentiment analysis datasets and their characteristics.

[†]Figure refers to a subset of the full corpus; the complete DynaSent dataset (Rounds 1 and 2) contains approximately 121k sentences.

2.2.2 Sentiment Analysis Models

The early approaches to sentiment analysis primarily employed lexicon and rule-based methods. These methods involved keyword matching techniques using manually or semi-automatically curated sentiment lexicons, such as SentiWordNet (Esuli and Sebastiani 2006) and LIWC (Pennebaker et al. 2015). Typically, sentiment scores were assigned to individual words or phrases, and the overall sentiment of a sentence or document was computed by aggregating these scores (Taboada et al. 2011). Rule-based methods extended these lexicons by introducing explicit linguistic rules or heuristics, designed to capture negation, intensification, and simple contextual clues (Hu and Liu 2004). However, these methods faced significant limitations due to their inability to capture context-dependent meanings or nuanced emotional expressions such as sarcasm, irony, or implicit sentiment, restricting their applicability to complex real-world datasets (Pang and Lee 2008; Liu 2012). Addressing the limitations of lexicon-based approaches, supervised machine learning models gained prominence in the early 2000s by leveraging annotated corpora to

learn sentiment patterns directly from data. Classical algorithms such as Support Vector Machines (SVM), Naive Bayes classifiers, and logistic regression proved particularly effective for sentiment classification tasks. Pang, Lee, and Vaithyanathan (2002) demonstrated that SVM-based classifiers significantly outperformed lexicon-based methods on binary sentiment classification benchmarks, highlighting the advantages of statistical learning approaches in adapting to linguistic variability.

Despite these advances, classical machine learning models introduced a notable bottleneck: their performance depended heavily on extensive manual feature engineering. Designing effective representations often required domain-specific knowledge and substantial experimentation with textual features such as uni-grams, bi-grams, part-of-speech tags, and syntactic dependency relations (Mohammad, Kiritchenko, and Zhu 2013). This reliance on handcrafted features ultimately limited the scalability and generalization of classical machine learning approaches across domains and linguistic contexts.

The advent of deep neural networks, including convolutional and recurrent architectures, marked a substantial shift in sentiment analysis modeling. CNN-based architectures introduced efficient feature extraction through convolutional operations, enabling models to automatically learn discriminative textual representations from raw input and substantially reducing the need for manual feature engineering (Kim 2014). In parallel, RNN-based architectures, particularly Long Short-Term Memory (LSTM) networks, enhanced the ability to model sequential dependencies and contextual information, capturing long-range interactions that are crucial for sentiment interpretation in longer texts and narratives (Hochreiter and Schmidhuber 1997; Tang et al. 2016).

A major breakthrough in sentiment analysis followed the introduction of Transformer-based architectures (Vaswani et al. 2017), which replaced recurrence with self-attention and enabled more effective modeling of contextual dependencies in text. Building on this architecture, the development of large pre-trained language models, including BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019) and the GPT family (Radford et al. 2018; Radford et al. 2019), substantially transformed sentiment analysis by providing transferable contextual representations that can be adapted to downstream tasks.

Beyond standard polarity classification, recent work has addressed two challenges that are particularly relevant for figurative language and conversational settings. Carvalho, Oliveira, and Silva (2023) systematically evaluated the impact of context in dialogue sentiment analysis, showing that including previous utterances (one or two turns) improves BERT-based classifiers, especially for negative sentiment detection in human–machine Portuguese conversations.

In practice, these models are typically fine-tuned on task-specific sentiment datasets, often yielding strong performance with comparatively limited task-specific labeled data relative to earlier feature-engineering pipelines (Devlin et al. 2019; Liu et al. 2019). In addition, large-scale generative models have demonstrated competitive task performance in zero-shot and few-shot settings via prompting, reducing the need for explicit supervised training in some scenarios (Brown et al. 2020). As a result, state-of-the-art sentiment analysis systems now routinely leverage pre-trained Transformers, achieving high benchmark performance while substantially reducing the need for manual feature engineering.

2.2.3 Evaluation Metrics

Evaluating sentiment analysis models has traditionally relied on standard classification metrics, accuracy, precision, recall, and F1 score, each capturing different aspects of predictive performance (Jurafsky and Martin 2019). Accuracy measures the proportion of correct predictions overall, but can be misleading in class-imbalanced datasets, where a model that predominantly predicts the majority class may still achieve high scores. Precision and recall address this by focusing respectively on the correctness of positive predictions and the coverage of true positive instances. The F1 score, their harmonic mean, provides a more balanced single-number summary and is particularly informative when sentiment classes are unevenly distributed. In multi-class settings, such as fine-grained five-class annotation schemes, a further distinction arises between macro-averaged F1, which weights each class equally and is sensitive to performance on minority classes, and micro-averaged F1, which aggregates counts across classes and favors the majority. The choice between these variants has meaningful consequences for how model performance is interpreted across different SA benchmarks.

Beyond these standard measures, more specialized metrics have been adopted to address limitations inherent in conventional evaluation. The Matthews Correlation Coefficient (MCC) is particularly well-suited to imbalanced datasets, combining all four cells of the confusion matrix into a single robust score that is sensitive to both classes regardless of their relative frequency (Chicco and Jurman 2020). Cohen's Kappa has been used as an inter-annotator reliability measure, correcting for chance agreement and thereby providing insight into both annotation quality and the difficulty of the sentiment classification task itself (Cohen 1960; Artstein and Poesio 2008).

Performance Benchmarks

The performance of sentiment analysis models has improved substantially over the past two decades, tracing the field's broader methodological transitions. Lexicon-based approaches, constrained by their inability to model context, typically achieved accuracies of around 60–70% on binary polarity benchmarks (Pang and Lee 2008). Supervised machine learning methods, notably Support Vector Machines and Naive Bayes classifiers, raised this ceiling to approximately 75–80% on the same tasks (Pang, Lee, and Vaithyanathan 2002; Pang and Lee 2004). The introduction of deep neural architectures, particularly convolutional and recurrent models, consistently surpassed classical methods and achieved binary classification accuracies above 85% on benchmarks such as SST-2 (Kim 2014; Tang et al. 2016). Fine-grained tasks, however, proved considerably more resistant to improvement: five-class schemes such as SST-5 typically yielded accuracies in the range of 45–55% for these architectures, reflecting the fundamental difficulty of distinguishing between adjacent sentiment categories (Socher et al. 2013).

The introduction of Transformer-based architectures (Vaswani et al. 2017), together with large-scale pre-trained models such as BERT (Devlin et al. 2019) and RoBERTa (Liu et al. 2019), produced substantial gains across both binary and fine-grained tasks. On binary benchmarks such as IMDb and SST-2, fine-tuned Transformer models routinely surpass 90–95% accuracy, establishing

a strong empirical ceiling for standard polarity classification. On fine-grained benchmarks such as SST-5, the gains are real but more modest: fine-tuned BERT and RoBERTa variants reach approximately 55–60% accuracy, while more advanced architectures such as DeBERTa-v3 achieve results in the 57–60% range, substantially above deep learning baselines, yet still reflecting the inherent difficulty of the five-class task (Devlin et al. 2019; Liu et al. 2019; Potts et al. 2021).

The current frontier is shaped by the emergence of large generative language models operating through prompting rather than supervised fine-tuning. Benchmarking studies have found that GPT-3.5, GPT-4, and comparable models, operating zero-shot, can match or in some cases exceed fine-tuned transfer learning baselines on binary and three-class sentiment tasks across diverse datasets (Krugmann and Hartmann 2024).

Era	**Model Type**	**Dataset**	**Accuracy**
Early 2000s	Lexicon-based	Movie reviews	∼60–70%
Mid 2000s	SVM / Naive Bayes	Movie reviews	∼75–80%
Mid 2010s	CNN / LSTM	SST-2	∼85%
Mid 2010s	CNN / LSTM	SST-5	∼45–55%
Late 2010s	BERT / RoBERTa	SST-2 / IMDb	>90–95%
2020s	DeBERTa / RoBERTa	SST-5	∼57–60%
2020s	Instr.-tuned LM[†]	ABSA (SemEval)	∼75–82% F1
2020s	Generative LLM (0-shot)	Binary (20 datasets)[‡]	∼80–95%[‡]

Table 2.4: Sentiment classification performance across model paradigms.

[†]Instr.-tuned LM: instruction-tuned language model. F1 range reflects reported ABSA results.
[‡]Based on Krugmann and Hartmann (2024), who benchmark GPT-3.5, GPT-4, and Llama 2 zero-shot across 20 datasets (>3,900 documents). Performance varies substantially by data origin: higher on structured product reviews, lower on Twitter data.

2.3 Speech Emotion Recognition

Speech Emotion Recognition (SER) is a subfield of speech processing that focuses on identifying and modeling affective states conveyed through vocal cues in the speech signal. SER systems typically classify emotions using acoustic and prosodic features (Schuller et al. 2011; El Ayadi, Kamel, and Karray 2011). Historically, SER gained traction in the late 1990s and early 2000s, driven by advances in signal processing, machine learning, and the growing availability of emotionally annotated speech corpora. The increasing ubiquity of voice-based technologies—from telephony systems to conversational interfaces—further accelerated interest in the field, since speech provides a rich and natural channel through which affective information is conveyed in human communication (Cowie et al. 2001).

The importance of SER has expanded considerably, with applications in fields such as customer service automation, human–robot interaction, and other affect-aware interactive systems, as well as emerging uses in mental health monitoring and adaptive learning (Schuller and Batliner 2014; Lee and Narayanan 2005). In the domain of media art and affect-driven immersive systems, SER has become an increasingly relevant mechanism for interactive installations and emotionally responsive environments. One of the earliest large-scale initiatives in this space was the EU Integrated Project *CALLAS* (approximately 2006–2010), which investigated multimodal affective interfaces for digital arts and entertainment ("EU project develops affective interface technologies for new media" 2007; Vogt et al. 2009). CALLAS developed interactive installations capable of interpreting vocal emotion, keyword input, and visual cues to modulate generative audiovisual systems in real time. A representative example is *E-Tree*, an augmented-reality installation in which emotional speech recognition, keyword spotting, and video analysis were combined to dynamically alter a virtual tree's growth, color, and behavior according to the user's affective state (Gilroy et al. 2008). Through such projects, CALLAS demonstrated how multimodal affect analysis could serve as a generative driver for interactive narrative and visual transformation. In parallel, the EU-funded *IRIS* Network of Excellence further consolidated research on interactive storytelling and affect-aware narrative systems, exploring how emotional input could shape character behaviour and narrative progression (Cavazza et al. 2008).

Recent research has begun to explore SER directly within applied VR scenarios. Saufnay, Etienne, et al. (2025) present a system specifically designed for virtual reality public speaking training, using acoustic features to classify ten discrete emotional states and deliver real-time feedback within immersive simulations. Zhang (2025) investigate the visualisation of poetry recitation emotions in VR, mapping detected vocal affect onto dynamic visual representations that externalize the expressive nuances of spoken performance. Together, these works illustrate the expanding role of SER in VR contexts, not only for artistic experimentation, but also for training, visualization, and embodied affective feedback.

2.3.1 Speech Emotion Recognition Databases

Early research in SER primarily relied on acted emotional speech datasets designed for discrete emotion classification, focusing on a limited set of prototypical emotions. One of the most influential datasets in this domain, the Berlin Emotional Speech Database (Emo-DB) (Burkhardt et al. 2005), features professional actors delivering emotionally expressive utterances in German and has served as a benchmark for early algorithmic evaluations. More recently, datasets such as the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (Livingstone and Russo 2018) have extended this paradigm by offering high-quality, standardized audio-visual recordings across both speech and song while retaining an acted and discrete emotion framework. The standardized production conditions of RAVDESS enabled more reliable cross-study comparisons than earlier corpora recorded under varied studio setups.

Following these early efforts, the Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) (Busso et al. 2008) marked a significant step forward by combining scripted and improvised dyadic dialogues with rich multimodal recordings. The dataset includes audio, video, and motion capture signals, enabling the study of emotional expression across multiple channels and supporting both categorical emotion labels and dimensional affect representations. Subsequently, datasets such as MSP-IMPROV (Busso et al. 2017) and MSP-PODCAST (Lotfian and Busso 2019) further advanced the field by emphasising increasingly naturalistic emotional speech. MSP-IMPROV bridges the gap between controlled acted corpora and fully spontaneous data by capturing semi-spontaneous emotional interactions through improvised dyadic scenarios, while MSP-PODCAST derives emotionally expressive speech from real-world podcast recordings, providing naturalistic data at a scale previously unavailable to the field.

Efforts toward culturally diverse data have also shaped SER research. Corpora such as RECOLA (Ringeval et al. 2013) provide naturalistic remote dyadic interactions with continuous 2D affect annotations (valence and arousal), supporting the study of emotion as it unfolds dynamically in realistic conversational settings. Complementing this, cross-cultural resources such as the SEWA database (Kossaifi et al. 2019) include participants from multiple cultural backgrounds and provide richly annotated audio-visual recordings, enabling the development of affect recognition systems that generalize beyond culturally homogeneous training data.

Recent dataset development has increasingly targeted subtle affective phenomena and interactional complexity. The SEMAINE database (McKeown et al. 2012) emphasises emotionally coloured, real-time conversational exchanges in a human–agent interaction setting, supporting the study of affect as it unfolds dynamically in dialogue. Complementing this line of work, MuSE (Jaiswal et al. 2020) was introduced as a multimodal dataset designed to elicit complex affective responses under stress, enabling analysis of emotion beyond clean, prototypical expressions and making it one of the few corpora to foreground stress-conditioned affective variability. Furthermore, corpora targeting non-literal and affectively ambivalent communication have gained prominence in multimodal affect research. The MUStARD dataset (Castro, Hazarika, Pérez-Rosas, et al. 2019), for instance, provides audiovisual utterances from television dialogues annotated for

sarcasm and accompanied by contextual turns, supporting research on the paralinguistic and contextual cues that shape sarcastic and ironic meaning.

Dataset	**Year**	**Speech Type**	**Annotation**	**Modality**
Emo-DB	2005	Acted	Categorical (7)	Audio
IEMOCAP	2008	Acted / Improv.	Categorical + Dimensional	Audio / Video / MoCap
RECOLA	2013	Spontaneous	Dimensional (2D)	Audio / Video / Physio
SEMAINE	2012	Elicited	Dimensional	Audio / Video
RAVDESS	2018	Acted	Categorical (8)	Audio / Video
MSP-IMPROV	2017	Acted (improv.)	Categorical (4)	Audio / Video
SEWA	2019	Spontaneous	Dimensional	Audio / Video
MSP-PODCAST	2019	Naturalistic	Categorical + Dimensional	Audio
MUStARD	2019	Naturalistic	Sarcasm (binary)	Audio / Video / Text
MuSE	2020	Elicited	Categorical + Dimensional	Audio / Video / Physio

Table 2.5: Representative Speech Emotion Recognition datasets and their characteristics.

Physio = physiological signals; Improv. = improvised; MoCap = motion capture.

2.3.2 Speech Emotion Recognition Models

Early approaches to SER primarily relied on rule-based and statistical methods grounded in acoustic signal processing. These systems employed hand-engineered features derived from prosodic, spectral, and voice quality characteristics, including fundamental frequency (F0), energy, speech rate, formant frequencies, and Mel-Frequency Cepstral Coefficients (MFCCs) (Ververidis and Kotropoulos 2006). Emotion classification was typically performed using simple pattern recognition techniques, such as distance-based classifiers, template matching, or early statistical models, where emotional states were inferred based on proximity to prototypical acoustic profiles. Although these approaches were intuitive and interpretable, they suffered from limited generalization, strong speaker dependence, and degraded performance in spontaneous or noisy conditions, restricting their applicability in real-world scenarios.

To overcome these limitations, classical machine learning methods became widely adopted for SER, leveraging annotated emotional speech corpora to train supervised classifiers. Techniques such as Support Vector Machines (SVMs), Gaussian Mixture Models (GMMs), k-Nearest Neighbours (k-NN), Hidden Markov Models (HMMs), and Random Forests demonstrated significant improvements over rule-based methods (Schuller, Steidl, and Batliner 2009). These models could better account for inter-speaker variability and emotion dynamics by learning statistical patterns from data. However, their performance remained highly dependent on handcrafted acoustic features and pre-processing pipelines, requiring domain expertise and extensive tuning. The need for robust and scalable feature representations posed a major bottleneck in extending these approaches to more diverse or real-time applications.

The advent of deep learning marked a transformative shift in SER by enabling models to learn hierarchical representations directly from raw or minimally processed speech signals. Convolutional Neural Networks (CNNs) proved effective at extracting local time–frequency patterns from spectrogram-based representations, while Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) architectures, excelled at modeling temporal dependencies and the evolution of emotional features across speech segments. The key limitation of RNNs in this context is their sequential processing of the input: hidden states are propagated one frame at a time, making it difficult to capture dependencies between distant parts of an utterance and causing gradients to vanish over long sequences. Trigeorgis et al. (2016) demonstrated an end-to-end SER framework that learns emotional representations directly from the raw waveform, substantially reducing reliance on hand-crafted features. Mirsamadi, Barsoum, and Zhang (2017) subsequently employed RNNs augmented with attention mechanisms to aggregate frame-level acoustic descriptors into utterance-level emotion predictions, showing that selectively weighting temporal frames according to their affective salience yielded consistent gains over uniform pooling. Deep architectures significantly reduced dependence on manual feature engineering and consistently outperformed classical SER methods on benchmark datasets; nevertheless, challenges remained in modelling long-range dependencies and in achieving robustness to speaker variability, linguistic diversity, and contextual factors present in naturalistic emotional speech.

These limitations motivated the adoption of attention-based and Transformer architectures (Vaswani et al. 2017), in which self-attention computes pairwise relationships between all frames in parallel, capturing global prosodic structure without the sequential bottleneck of recurrence. Pre-trained self-supervised models such as wav2vec 2.0 (Baevski et al. 2020) and HuBERT (Hsu et al. 2021) extended this paradigm by leveraging large quantities of unlabelled audio to learn transferable acoustic representations that can subsequently be fine-tuned for emotion recognition. The two models differ in their pre-training objectives: wav2vec 2.0 optimises a contrastive loss over quantised latent speech units, whereas HuBERT predicts discrete cluster assignments produced by an offline k-means step applied to acoustic features. In both cases, fine-tuning for SER typically discards the automatic speech recognition head, replaces it with a lightweight emotion classifier, and either freezes the lower transformer layers or applies a small learning rate to the full encoder. More recently, Ma et al. (2023) introduced emotion2vec, a self-supervised model pre-trained explicitly on emotional speech data rather than repurposed from ASR objectives, establishing a new reference point for SSL-based SER. In parallel, large supervised models such as Whisper (Radford et al. 2023), trained on extensive labeled audio–text corpora, have demonstrated that their encoder representations can be repurposed for affective and paralinguistic analysis without task-specific pre-training. Moreover, multimodal Transformer architectures have enabled effective fusion of speech with complementary modalities, including facial expressions, body gestures, and textual transcriptions. Large multimodal pretraining frameworks such as VATT (Akbari et al. 2021) (Video-Audio-Text Transformer) learn shared representations from raw audio, video, and text via self-supervised contrastive objectives. Although not emotion-specific, these pretrained multimodal embeddings can be adapted to affective computing tasks, supporting robust

multimodal emotion recognition in complex interactive environments. Multimodal Transformers are particularly valuable for addressing affective ambiguity and mixed emotional states, and for modeling subtle pragmatic phenomena, such as sarcasm and ironic prosody, that often emerge in spontaneous, context-dependent communication.

Today, state-of-the-art SER systems regularly combine pre-trained Transformer-based audio encoders with lightweight downstream emotion classification layers, enabling transfer learning across tasks and domains with minimal labeled data. This paradigm shift has substantially reduced the overhead of feature engineering and annotation, and has opened new directions at the intersection of large language models and speech, where researchers are exploring whether LLM-based systems can reason over affective speech in zero- or few-shot settings.

2.3.3 Evaluation Metrics

Evaluating SER models traditionally relies on classification metrics including accuracy, precision, recall, and the F1 score (Schuller, Steidl, and Batliner 2009). Accuracy provides a basic measure of overall correctness but can be misleading in the presence of class imbalance—a common characteristic of emotional speech corpora where certain classes (e.g., neutral) tend to dominate. Precision and recall therefore offer more informative, class-specific insights, particularly for underrepresented or acoustically ambiguous emotions. The F1 score, as the harmonic mean of precision and recall, remains a widely adopted summary metric in multi-class emotion classification settings. In multi-class SER tasks, a further distinction arises between macro-averaged F1, which weights each emotion class equally and is therefore sensitive to performance on minority classes, and micro-averaged F1, which aggregates counts across all classes and tends to favour the majority. This distinction has meaningful consequences for how model performance is interpreted across datasets with different class distributions and sizes.

To address class imbalance more explicitly, Unweighted Average Recall (UAR) has become a standard evaluation metric in SER research, particularly in academic benchmarks and challenge evaluations (Eyben et al. 2016). UAR computes recall independently for each class and averages the results without weighting by class frequency, making it well suited to skewed emotion distributions. Confusion matrices further complement these metrics by providing detailed insights into inter-class confusions, which are especially relevant when distinguishing between emotions with subtle acoustic differences, such as sadness versus neutrality or anger versus fear.

Beyond standard classification measures, inter-annotator agreement metrics have also been employed in SER evaluation, reflecting the inherently subjective nature of emotional labelling. Cohen's Kappa is used to assess annotation reliability, correcting for chance agreement and providing insight into both the quality of the ground-truth labels and the intrinsic difficulty of the emotion classification task (Cohen 1960; Artstein and Poesio 2008). Low inter-annotator agreement in a corpus signals that the emotional categories themselves may be perceptually ambiguous, which places an upper bound on model performance that is often overlooked in purely quantitative benchmarking.

In addition to categorical evaluation, SER systems trained for continuous affect prediction are commonly assessed using regression-based metrics such as Pearson's correlation coefficient (PCC) and the Concordance Correlation Coefficient (CCC). PCC measures the linear association between predicted and ground-truth affect trajectories, while CCC additionally penalises systematic offsets in mean or scale, making it more sensitive to prediction bias and better suited to evaluating continuous affect estimation in practice. These measures are particularly relevant for datasets with dimensional emotion annotations—such as RECOLA (Ringeval et al. 2013) and SEWA (Kossaifi et al. 2019)—where models aim to predict time-varying trajectories of affective dimensions such as arousal and valence rather than discrete emotion labels.

Performance Benchmarks

The historical progression in SER reflects a clear upward trajectory that closely mirrors the broader methodological transitions described in the preceding sections. Performance has improved from around 60–70% with early rule-based and classical machine learning approaches to above 80% with contemporary deep learning and pre-trained Transformer-based models, particularly on well-annotated benchmark datasets. Rule-based and GMM-based systems applied to controlled acted corpora such as Emo-DB typically achieved classification accuracies in the 60–70% range (Schuller, Steidl, and Batliner 2009). Supervised machine learning methods, notably SVM and HMM classifiers operating on handcrafted MFCC and prosodic features, raised this ceiling to approximately 70–80% on acted datasets and to roughly 60–65% UAR on more challenging naturalistic corpora such as IEMOCAP (Schuller, Steidl, and Batliner 2009).

Era	Model Type	Dataset	Metric	Score
Early 2000s	Rule-based / GMM	Emo-DB	Acc.	∼60–70%
Mid 2000s	SVM + handcrafted features	Emo-DB	Acc.	∼70–80%
Early 2010s	SVM / HMM + MFCC	IEMOCAP	UAR	∼60–65%
Mid 2010s	CNN / LSTM (end-to-end)	IEMOCAP	UAR	∼65–72%
Mid 2010s	CNN / ResNet (log-Mel / MFCC)	RAVDESS	Acc.	∼70–80%
Late 2010s	RNN + attention	IEMOCAP	UAR	∼68–75%
2020s	wav2vec 2.0 (fine-tuned)†	IEMOCAP	UAR	∼72%
2020s	wav2vec 2.0 / HuBERT (fine-tuned)	IEMOCAP	WA	∼75–80%
2020s	wav2vec 2.0 / HuBERT (fine-tuned)	RAVDESS	UAR	∼84%

Table 2.6: Speech Emotion Recognition performance across model paradigms.

Performance varies substantially depending on evaluation protocol (speaker-dependent vs. speaker-independent), number of emotion classes, and dataset subset used. All figures are approximate and should be interpreted as indicative of paradigm-level trends rather than definitive state-of-the-art results.

UAR = Unweighted Average Recall; WA = Weighted Accuracy; Acc. = standard classification accuracy.

†Based on Pepino, Riera, and Ferrer (2021): wav2vec 2.0 embeddings with shallow downstream classifier, audio-only condition, IEMOCAP improvised sessions only.

The introduction of deep learning architectures, including end-to-end CNN–LSTM networks trained directly on spectrograms or raw waveforms, produced consistent gains across both acted and semi-spontaneous datasets. On IEMOCAP, CNN and LSTM-based models achieved UAR

values in the range of 65–72%, while on RAVDESS, CNN and ResNet architectures trained on log-Mel spectrograms or MFCC representations reached classification accuracies of approximately 70–80%, demonstrating the effectiveness of deep feature learning even on acted emotional data (Issa, Demirci, and Yazici 2020).

The adoption of pre-trained Transformer-based speech encoders marked the most substantial leap in SER performance. Self-supervised models such as wav2vec 2.0 (Baevski et al. 2020) and HuBERT (Hsu et al. 2021), fine-tuned on downstream emotion recognition tasks, substantially outperformed all prior approaches. On IEMOCAP, wav2vec 2.0-based systems achieved UAR values of approximately 72% in audio-only configurations (Pepino, Riera, and Ferrer 2021), while fine-tuned HuBERT models reached weighted accuracies of 75–80% under speaker-dependent evaluation protocols (Wang, Boumadane, and Heba 2022). On RAVDESS, fine-tuned self-supervised models similarly advanced UAR to approximately 84% (Pepino, Riera, and Ferrer 2021). It should be noted that performance varies substantially depending on evaluation protocol (speaker-dependent vs. speaker-independent), number of emotion classes, and dataset characteristics—including whether speech is acted or spontaneous—making direct cross-study comparisons difficult and reported figures indicative rather than definitive.

2.4 Bimodal Speech Emotion Analysis

Multimodal analysis of speech reveals a fundamental property of affective communication: emotional meaning is not necessarily coherent across modalities. Specifically, the emotional valence of linguistic–semantic content and vocal–prosodic expression can exhibit systematic divergences. These incongruities constitute meaningful communicative strategies, most notably in phenomena such as irony and sarcasm.

Early empirical work sought to quantify the relative contribution of each channel to emotional judgment. Davitz (1964) reported experiments that attempted to isolate intonational cues from verbal meaning by progressively filtering speech to reduce intelligibility, estimating the degree to which affect is attributed to vocal tone when semantic content is rendered unavailable. A complementary paradigm employed highly constrained utterances, such as recited letters or numbers, produced with varied affective intonations, examining how listeners map prosodic form onto emotional interpretation in the absence of meaningful propositional content.

Mehrabian and Wiener (1967) raised methodological objections to both approaches: filtering degrades rather than cleanly removes semantic information, and artificially constrained verbal material may not support inferences that generalize to natural speech. To address these concerns, they designed experiments using deliberately inconsistent communications, in which affect was judged from a small set of short, prerecorded utterances delivered in systematically varied tones of voice. Extending this line of inquiry to the visual modality, Mehrabian and Ferris (1967) examined how listeners integrate auditory and facial cues when inferring speaker attitudes. Participants heard a female speaker utter the word *maybe* in three tones intended to convey liking, neutrality, or disliking, while simultaneously viewing photographs of female faces expressing corresponding affective states. Facial information proved the stronger predictor of judged attitude, though the authors cautioned explicitly against generalizing these channel weights beyond comparably constrained communicative situations.

The question of cross-modal integration resurfaces in computational approaches to multimodal affect modeling, recast as an architectural design problem: how should information from multiple modalities be combined? Historically, the field focused on two paradigms: Feature-level fusion, in which modality-specific descriptors are concatenated into a single joint representation, offers simplicity but can introduce high-dimensional feature spaces and elevated overfitting risk (Atrey et al. 2010). Decision-level fusion, by contrast, trains separate classifiers per modality and combines their outputs, preserving modality-specific inductive biases while allowing the system to exploit complementary cues flexibly (Kittler et al. 1998). More recently, intermediate fusion and attention-based mechanisms have emerged, allowing for the learning of complex cross-modal correlations within the hidden layers of neural architectures.

2.4.1 Irony as Semantic–Prosodic Opposition

Understanding language often requires inferring meanings that go beyond propositional content, integrating pragmatic cues, contextual information, and speaker intentions (Gibbs and Colston

2012; Gibbs 2000). Figurative (or non-literal) utterances rely on this inferential process, as their interpretation emerges from the interaction between what is said, how it is expressed, and the communicative situation in which it occurs (Gibbs and Colston 2012). Roberts and Kreuz (1994) identify several common forms of figurative language in everyday communication, including hyperbole, idiom, indirect request, irony, understatement, metaphor, rhetorical question, and simile.

The concept of irony has been historically associated with dissimulation or pretense (Szenberg and Ramrattan 2014; Knox 1961). In classical rhetoric, Quintilian described irony as a figure of speech in which the intended meaning contradicts the literal formulation of the utterance (Quintilian 1921). From a pragmatic perspective, Grice (1975) framed irony as a case of conversational implicature, whereby speakers deliberately violate conversational maxims to convey a meaning that differs and often opposes the literal content. Relevance Theory further conceptualizes irony as an echoic or attributive use of language that expresses a speaker's evaluative attitude toward a thought or proposition (Wilson and Sperber 2012). These approaches highlight irony's inferential nature, its intrinsic ambiguity, and its dependence on contextual and prosodic cues (A. Braun and Anita Schmiedel 2018).

Common classifications distinguish several forms of irony, including verbal, dramatic, cosmic, and romantic irony (Muecke 1969; Colebrook 2003). More fine-grained analyses propose subtypes of verbal irony, such as propositional, embedded, like-prefixed, and illocutionary irony, each differing in their reliance on contextual inference, linguistic incongruity, and prosodic marking (Camp 2012).

Illocutionary irony (or illocutionary sarcasm) refers to cases in which ironic meaning is not primarily encoded in the lexical content of the utterance, but rather emerges through its delivery across non-verbal modalities. In such instances, the propositional content may appear fully sincere at the textual level, while prosodic, gestural, or visual cues convey an opposing evaluative stance. An "ironic tone of voice" has been clearly identified in more exaggerated or "dripping" cases, where prosodic modulation strongly signals the speaker's intent (Bryant and Fox Tree 2005). However, this distinction becomes considerably more challenging in spontaneous speech, where contextual information may be limited and prosodic cues are more subtle or variable, making ironic intent harder to detect (Bryant and Fox Tree 2005).

Irony frequently overlaps with sarcasm, although several authors distinguish between sarcastic irony ("praise by blame") and so-called kind irony ("blame by praise") (Anolli, Ciceri, and Infantino 2000, 2002). Sarcasm is often characterized by a more overt critical stance and may function as an indirect form of evaluation or face management within social interaction (Brown and Levinson 1987; Jorgensen 1996). Kind irony, by contrast, is less common and tends to be described as more affiliative or empathetic, despite retaining its non-literal structure (Anolli, Ciceri, and Infantino 2002).

A growing body of research shows that both irony and sarcasm can be distinguished through systematic analysis of acoustic patterns, such as modulations in pitch, intensity, tempo, and voice quality, that partially overlap with those observed in emotional speech (Fónagy 1971; Cheang and Pell 2008; Mauchand et al. 2021). This convergence suggests that speech emotion recogni-

tion resources and multimodal affective datasets may provide a valuable methodological basis for studying emotional divergences and prosodic ironies.

2.4.2 Acoustic and Prosodic Markers of Irony

Several researchers have studied prosodic features as a key attribute of irony (see Table 1). Quintilian mentions that irony is recognizable by the speaker's tone and body language (Quintilian 1921). In the Pretense Theory of Irony, Clark and Gerrig (1984) described the ironic tone of voice as a mechanic where speakers often exaggerate the tone to make their pretense evident to the initiated audience. Cutler (1974) argued that posed irony could be recognized through nasalization (the entire sentence or a portion of it may be nasalized), slowed speech (the speaking rate may be reduced), and exaggerated stress (i.e., one or more words may receive exaggerated emphasis, particularly by exaggeratedly lengthening stressed syllables). Other perceptual approaches in English have been proposed by Haverkate (1990) and Kreuz and Roberts (1995). Rockwell (2000) shifts attention to how prosodic features,such as pitch, tempo, and intensity, serve as cues for Sarcasm. He conducted an experiment where speakers were recorded reading sentences under three conditions: non-sarcasm, spontaneous sarcasm, and posed sarcasm. These recordings were then filtered to remove the semantic content, allowing listeners to focus solely on vocal features. The results showed that listeners could reliably distinguish posed sarcasm from non-sarcasm but had difficulty distinguishing it from spontaneous sarcasm from non-sarcasm. Rockwell's results showed that a slower speech tempo, a louder intensity, and a lower pitch marked Sarcasm .

Bryant and Fox Tree (2002) used ironic and non-ironic utterances from radio talk shows, presenting them either in auditory or written form, with or without contextual information. Textually ambiguous ironic utterances were rated as equally sarcastic as non-ironic ones when read silently, but were judged as significantly more sarcastic when heard, suggesting that prosodic cues were driving identification. Extending this work, Bryant and Fox Tree (2005) conducted acoustic analyses and filtering experiments on a broader set of utterances, distinguishing between *dry* sarcasm (textually ambiguous) and *dripping* sarcasm (prosodically exaggerated). Acoustic measurements revealed no reliable global differences between dry ironic and non-ironic speech across any of the five parameters examined. For dripping utterances, the only statistically significant acoustic difference was lower amplitude variability in ironic speech; a marginal difference in mean F0 was also observed. Crucially, when listeners rated content-filtered versions of dripping utterances, ironic items were perceived not only as more sarcastic but also as angrier and more inquisitive, suggesting that listeners do not draw on a dedicated ironic vocal signal but instead integrate overlapping affective and prosodic cues in their interpretation.

Attardo et al. (2003) used multimodal stimuli from television sitcoms and found no singular ironic intonation pattern; instead, pitch appeared to function as a contrastive marker, where variations from expected pitch levels, rather than any fixed tonal configuration, signal the presence of irony. Cheang and Pell (2008) compared posed sarcastic speech against three other attitudes (humour, sincerity, and neutrality) across a comprehensive set of acoustic parameters. The most robust finding was a significant reduction in mean fundamental frequency (F0) in sarcastic utterances

	Dry		*Dripping*	
Acoustic Variable	**Ironic**	**Non-ironic**	**Ironic**	**Non-ironic**
Mean F0 (semitones)	12.8	11.6	15.3	11.2†
F0 Range (semitones)	32.6	26.5	25.7	32.7
F0 Variability (semitones)	3.37	4.03	2.64	2.69
Amplitude Variability (dB SD)	11.36	11.93	14.5	16.1*
Mean Syllabic Duration (ms)	187	201	233	218

Table 2.7: Acoustic analysis of dry and dripping ironic versus non-ironic utterances (Bryant and Fox Tree 2005)

Note. All F0 comparisons performed on semitone values. $^{\dagger}p < .10$; $^{*}p < .05$; all other differences non-significant. Adapted from Bryant and Fox Tree (2005), Tables 1 and 2.

relative to all other attitudes, irrespective of linguistic context, indicating that sarcastic speech is systematically produced at a lower pitch. Sarcasm was further distinguished from sincerity, though not from humour, by reductions in both F0 standard deviation and harmonics-to-noise ratio (HNR), suggesting that sarcastic speech tends to be less pitch-variable and noisier in vocal quality than sincere speech. Additional cues emerged in specific linguistic contexts only: for short keyphrase utterances, sarcasm was produced at a slower rate and with a more restricted F0 range, while resonance differences were observed primarily in longer sentence and combined sentence exemplars (see Table 2.8). These findings suggest that sarcasm is marked by a core cluster of prosodic cues, particularly lowered mean F0 and reduced HNR, supplemented by context-dependent features that interact with utterance length and linguistic form.

Acoustic Parameter	**Sarcasm vs. Humour**	**Sarcasm vs. Sincerity**	**Phrase Type**
Mean F0	Lower*	Lower*	All types
F0 standard deviation	ns	Lower*	All types
F0 range	ns	Lower*	Keyphrases
Mean amplitude	ns	ns	—
Amplitude range	ns	ns	—
Speech rate	Lower*	Lower*	Keyphrases
HNR	ns	Lower*	All types
Resonance (1/3 octave)	Different*	Different*	Sentences & combined

Table 2.8: Summary of acoustic cues distinguishing sarcasm from other attitudes (Cheang and Pell 2008)

Note. "Lower*" denotes a statistically significant reduction in sarcasm relative to the comparison attitude. "ns" = not significant. "Different*" denotes a significant difference in the one-third octave spectral profile. Adapted from Cheang and Pell (2008).

Burkhardt, Schmitt, and Schuller (2018) investigate the detection of vocal irony seeking discrepancy between textual sentiment and vocally expressed valence. Using a dataset of 937 utterances labeled for irony and anger by six listeners, the study explores the effectiveness of prosodic features such as pitch, tone, and speech rate as indicators of irony without relying on semantic

content. The researchers employ a Support Vector Machines approach for acoustic emotion classification. Two feature sets were utilized: the knowledge-driven Geneva Minimalistic Acoustic Parameter Set (GeMAPS) (Eyben et al. 2016), which captures statistical moments of prosodic and voice quality parameters, and the larger-scale computational paralinguistic challenge feature set (ComParE), which encompasses thousands of low-level descriptors ComParE (Eyben et al. 2013). Baseline performance achieved an unweighted average recall (UAR) of 69.3% for irony detection and 64.1% for anger.

There have been reported gender differences in prosodic features between male and female speakers (Anolli, Ciceri, and Infantino 2002; B. Braun and Anna Schmiedel 2018; Li, Poria, and Hazarika 2024). Tepperman, Traum, and Narayanan (2006) investigated sarcasm detection using contextual, prosodic, and acoustic features in spoken dialogue. The dataset comprised 131 occurrences of the phrase "yeah right" from the Switchboard and Fisher corpora, with 23% labeled as sarcastic. Contextual cues, such as laughter, gender, turn position, pauses, and conversational role (e.g., response to a question), were paired with prosodic features like pitch, energy, duration, and low-level acoustic features, including MFCCs. Results show that contextual and acoustic features significantly outperformed prosodic features, achieving up to 87% accuracy, while prosodic features alone yielded the lowest performance (69%). The study identified laughter as the most critical contextual feature and specific pitch and energy metrics as the most relevant prosodic indicators.Rakov and Rosenberg (2013) addressed the limitations of Tepperman, Traum, and Narayanan (2006) study by creating the Daria Sarcasm Corpus, a dataset of sarcastic and sincere utterances derived from an animated sitcom. This corpus was explicitly designed to mitigate the ambiguity and sparsity of sarcastic examples in the Switchboard and Fisher corpora, enabling more robust analysis. They proved to detect sarcasm with a significantly improved accuracy of 81.57% using a LogitBoost classifier. Unlike Tepperman's reliance on contextual cues, Rakov and Rosenberg demonstrated that word-level prosodic features, such as reduced pitch range and consistent intensity contours, could independently contribute to accurate sarcasm detection.

Castro, Hazarika, Poria, et al. (2019) introduced the MUStARD dataset consisting of 690 balanced samples of sarcastic and non-sarcastic utterances drawn from popular TV shows. Each sample includes text, audio, and video modalities, preceding conversational context, and speaker identifiers. They employed Support Vector Machines (SVM) with early fusion to combine multimodal features, including BERT-based text embeddings (Devlin et al. 2019), low-level acoustic features (MFCC, mel-spectrogram, and spectral centroid, extracted via Librosa), and visual frame representations obtained from a pretrained ResNet-152 (He et al. 2016). Their speaker-dependent experiments demonstrated that incorporating multimodal information significantly outperformed unimodal approaches, with the best bimodal setup (text and video) achieving an F1-score of 71.6% and a relative error rate reduction of up to 12.9% over the best unimodal baseline (see Table 2.9). However, performance gains were marginal in the speaker-independent setup, where the audio modality played a more prominent role and the tri-modal model achieved an F1-score of 62.8%.

Chauhan et al. (2020) extended the MUStARD dataset by incorporating categorical sentiment annotations (positive, negative, neutral) and emotion labels (anger, excitement, fear, sadness,

Setup	Modality	Precision	Recall	F1-Score
Speaker-dependent	T	65.1	64.6	64.6
	A	65.9	64.6	64.6
	V	68.1	67.4	67.4
	T+A	66.6	66.2	66.2
	T+V	**72.0**	**71.6**	**71.6**
	A+V	66.2	65.7	65.7
	T+A+V	71.9	71.4	71.5
Speaker-independent	T	60.9	59.6	59.8
	A	65.1	62.6	62.7
	V	54.9	53.4	53.6
	T+A	**64.7**	**62.9**	**63.1**
	T+V	62.2	61.5	61.7
	A+V	64.1	61.8	61.9
	T+A+V	64.3	62.6	62.8

Table 2.9: Sarcasm classification results on the MUStARD dataset (Castro, Hazarika, Poria, et al. 2019)

Note. T = text, A = audio, V = video. Bold denotes best-performing multimodal configuration per setup. All results are weighted F-scores across sarcastic and non-sarcastic classes. Adapted from Castro, Hazarika, Poria, et al. (2019), Tables 2 and 3.

surprise, frustration, happiness, neutral, and disgust). They proposed a multi-task deep learning framework that utilized these annotations alongside multimodal inputs to jointly analyze sarcasm, sentiment, and emotion. By introducing two attention mechanisms—Inter-segment Inter-modal Attention and Intra-segment Inter-modal Attention—they captured cross-modal and intra-modal relationships. Their experiments revealed that integrating sentiment and emotion analysis improved sarcasm detection performance, achieving up to 5% higher F1 scores than single-task models. Ray, Mishra, and Poria (2022) introduced MUStARD++, an extended version of the original MUStARD dataset. They doubled the size of the dataset, adding new sarcastic and non-sarcastic video samples from similar sitcom genres annotated with valence, arousal, and sarcasm-type metadata. Their work addressed errors in the emotion labels provided, identifying and correcting 343 mislabelings.

Ray, Mishra, and Poria (2022) conducted extensive experiments with multimodal fusion models, incorporating features from text (BART embeddings), audio (low-level descriptors such as MFCC and prosodic features), and video (ResNet-152 frame representations). Their collaborative gating-based fusion architecture outperformed prior approaches, achieving a weighted F1-score of 74.2% in sarcasm detection under speaker-dependent conditions on the original MUStARD dataset, with a 9.25% improvement over unimodal baselines. Li, Poria, and Hazarika (2024) used MUStARD++ to investigate how prosodic and semantic features interact to convey sarcasm. The utterances were classified into propositional sarcasm (333 instances), embedded sarcasm (87 instances), and illocutionary sarcasm (178 instances), each characterized by varying reliance on semantic and prosodic signals. Prosodic analysis at the utterance level revealed that sarcastic ex-

pressions were marked by a lower mean fundamental frequency (F0) and higher mean amplitude than neutral utterances. At the key phrase level, embedded Sarcasm showed minimal prosodic modulation. At the same time, illocutionary Sarcasm exhibited distinct patterns, including longer duration and increased amplitude. Additionally, propositional sarcasm phrases showed a reduced F0 range compared to neutral phrases.

More recently, Bhosale et al. (2023) addressed the imbalance in the sarcasm-type category of MUStARD++ by augmenting it with 90 sarcastic and 74 non-sarcastic clips drawn from the TV series *House MD*, producing MUStARD++ Balanced. Rigorous benchmarking with state-of-the-art language, speech, and visual encoders yielded a 2% macro-F1 improvement over the prior MUStARD++ benchmark, with a further 2.4% gain when training on the balanced extension. The trend in the field has since shifted towards leveraging large multimodal models. Saha et al. (2025) introduced MUStReason, a diagnostic benchmark built on MUStARD++ Balanced that enriches each instance with modality-specific perceptual cues and structured reasoning annotations. Rather than treating sarcasm detection as a binary classification problem, MUStReason disentangles it into a perception stage, identifying relevant cues in text, audio, and video, and a reasoning stage that infers implied intent from those cues.

2.4.3 Cross-Linguistic Prosodic Patterns of Irony

Ironic prosody has been investigated in different languages (Fónagy 1971; Anolli, Ciceri, and Infantino 2000, 2002; Scharrer, Christmann, and Knoll 2011; Niebuhr 2014; B. Braun and Anna Schmiedel 2018; Bolyanatz, Soto-Barba, and Guerrero 2024). Fónagy (1971) identifies several key features that contribute to the perception of Irony in Hungarian, including pitch contour, speech rate, voice quality, and intensity levels. Fónagy describes how Irony is often conveyed through a specific pitch contour and a deliberate slowing of speech rate, with a pace typically one-third of the standard rate. Anolli, Ciceri, and Infantino (2002) analyzed posed verbal irony (sarcasm and kind irony) in longer utterances in Italian, where they found that sarcasm was characterized by a significant increase in mean pitch (F0), with a broader pitch range and higher variability. In contrast, kind Irony exhibited a smaller increase in F0 and less pitch variability. Both forms of Irony showed elevated intensity levels, with sarcastic Irony demonstrating increases compared to normal speech. Sarcasm was also spoken at a slower rate than kind irony and normal speech. Niebuhr (2014) analyzed phonetic differences between sarcastic and neutral utterances in German. A production experiment was conducted using 20 sentences spoken by 10 participants in neutral and sarcastic tones. The results revealed that sarcastic utterances were characterized by longer durations, lower and flatter F0 contours, lower intensity, breathier voice quality, and higher segmental reduction.

B. Braun and Anna Schmiedel (2018) investigated the phonetic properties of verbal irony using short, single-word German utterances, distinguishing between sarcasm (*blame by praise*) and kind irony (*praise by blame*) and including neutral stimuli as a baseline. In the production experiment, sarcasm was marked by a significantly lower mean F0, reduced F0 standard deviation, and narrower F0 range relative to sincere praise, as well as longer utterance duration. Kind irony,

by contrast, exhibited the opposing pattern: higher F0 range than sincere blame, with a tendency toward greater pitch variability, though most kind irony F0 differences did not reach significance. Intensity patterns also diverged between irony types, with sarcastic utterances softer than sincere praise, and a reversed pattern for kind irony; however, intensity differences were further modulated by speaker gender, with female speakers marking irony through greater intensity relative to neutral, and male speakers through reduced intensity. The perception experiment, involving 44 listeners, yielded an overall recognition rate of approximately 70%, with the positive stimulus set (sarcasm) showing a higher overall recognition rate than the negative set (kind irony) (72% vs. 65%). Within each stimulus set, however, kind irony was identified more accurately than sarcasm (73% vs. 70%), while praise was the most reliably recognised category overall (77%). Speaker sex, but not listener sex, significantly influenced recognition accuracy, with female speakers' utterances decoded more reliably across all categories except blame (see Table 2.10).

	Positive set (Sarcasm)		**Negative set (Kind Irony)**	
Parameter	**Direction**	**Sig.**	**Direction**	**Sig.**
Production (ironic vs. sincere counterpart)				
Duration	Sarcasm > praise	*	Kind irony > blame	*
Mean F0	Sarcasm < praise	*	Kind irony > blame	ns
F0 SD	Sarcasm < praise	*	Kind irony > blame	ns
F0 range	Sarcasm < praise	*	Kind irony > blame	*
Intensity	Sarcasm < praise	*	Kind irony > blame	ns
Perception (recognition rates)				
Ironic utterance	Sarcasm: 70%		Kind irony: 73%	
Sincere utterance	Praise: 77%		Blame: 62%	
Overall	≈70%			

Table 2.10: Production findings and perception recognition rates for sarcasm and kind irony (B. Braun and Anna Schmiedel 2018)

Note. * = significant ($p < .05$ or better); ns = not significant. F0 SD and range expressed in semitones. Speaker sex, but not listener sex, significantly modulated recognition rates. Adapted from B. Braun and Anna Schmiedel (2018), Table 4.

Scharrer, Christmann, and Knoll (2011) conducted a study to explore prosodic features of Irony in German speech, focusing on how ironic criticism differs from literal speech in terms of voice modulations. Their analysis revealed that speech was characterized by a lower mean fundamental frequency (F0), higher energy levels, and increased vowel duration, in contrast to literal speech. While the overall F0 contours did not differ significantly between the two speech types, the tonal variations within an utterance remained similar. They describe vowel hyperarticulation in ironic speech, a feature that had not previously been considered an irony marker. Additionally, contrary to expectations, these voice modulations were found to occur independently of visual irony cues.González-Fuente, Prieto, and López (2016) conducted a detailed analysis of the acoustic cues involved in verbal irony recognition in French, focusing on the role of prosodic features such as pitch range expansion, syllable lengthening, and specific intonational contours. Results

revealed that the most effective cue for irony recognition was the combination of all three modified features, with duration and intonation emerging as particularly influential in signaling Irony. In contrast, pitch range expansion alone was less effective. Bolyanatz, Soto-Barba, and Guerrero (2024) investigated prosodic characteristics of irony in Chilean Spanish through sociolinguistic interviews with 15 women. Researchers analyzed 197 ironic utterances across three subtypes: jocularity, hyperbole, and rhetorical questions. Acoustic measures were compared with pre-ironic utterances, including pitch, duration, and voice quality. Results showed significant prosodic differences, such as increased pitch range in jocularity and hyperbole, creakier voice quality in rhetorical questions, and shorter vowel duration in hyperbole.

Rao, Ronquest, and Hall-Lew (2022) investigate how bilingual speakers of English and Spanish express sarcasm and sincerity prosodically, focusing on heritage Spanish speakers residing in the USA. The study analyzes data from 19 English-dominant, English-Spanish bilinguals, including a subset of participants from the same family. The analysis, which examines prosodic features such as f0 mean, f0 range, and speech rate using Praat, reveals that sincere speech is produced at a faster speech rate than sarcastic speech, with bilinguals speaking faster in English than in Spanish. The findings also indicate that sarcasm versus sincerity is influenced by age, gender, and language dominance, with age modifying the relationship between attitude and speech rate. Additionally, the study highlights that speaking in Spanish results in a higher f0 mean and range compared to English, and these prosodic features are shaped by complex interactions between language, attitude, and other sociolinguistic variables, including bilingual type and family speech patterns. Notably, the analysis of intrafamilial data demonstrates that parents and their adult children exhibit similar prosodic patterns in Spanish, suggesting the role of source input variety in shaping prosodic expression.

Table 2.11: Prosodic Irony Studies

Author & Year	Attitude	Database	Gender	Language	Features	Result
Rockwell (2000)	Sarcasm	Speech recordings of sentences in three conditions (nonsarcasm, spontaneous sarcasm, posed sarcasm)	Male and Female	English	Pitch (mean, variation); Amplitude (mean, variation); Time; Resonance; Articulation	Listeners distinguished posed sarcasm from non-sarcasm but not spontaneous sarcasm from non-sarcasm. Sarcasm is slower tempo, greater intensity and a lower pitch level.
Anolli et al. (2000)	Sarcasm, kind irony	Posed sarcasm	Male	Italian	Pitch (Mean, min, max, range, SD); Energy (Mean, range, min, max, SD); Time (Duration, Number of pauses, Pauses duration, Rate of articulation (syl/s))	Sarcastic irony has high pitch, constant energy, and bounded voice quality, while kind irony has moderate pitch, lower energy, and mild voice quality. Irony uses exaggerated suprasegmental features compared to normal speech.
Anolli et al. (2002)	Sarcasm, kind irony	Posed sarcasm	Male	Italian	Pitch (Mean, min, max, range, SD); Energy (Mean, range, min, max, SD); Time (Duration, Number of pauses, Pauses duration, Rate of articulation (syl/s))	Sarcastic irony has high pitch, constant energy, and bounded voice quality, while kind irony has moderate pitch, lower energy, and mild voice quality. Irony uses exaggerated suprasegmental features compared to normal speech.
Nakassis & Snedeker (2002)	Ironic insults (sarcasm) and ironic compliments	Speech recordings of story scenarios with children characters	Female	English	Intonation (positive/negative); Context discrepancy	Adults comprehended irony better than children. Both relied on contextual cues. Discordant utterances were harder for children to process, and intonation relationally impacted irony interpretation.
Bryant and Fox-Tree (2002)	Ironic and nonironic	Spontaneous speech recordings from radio talk shows	Male and Female	English	Amplitude variability; context	Irony rated as more sarcastic when context was included; listeners used various features rather than specific ones to identify irony.

Author & Year	Attitude	Database	Gender	Language	Features	Result
Attardo et al. (2003)	Irony	Multimodal stimuli from television situation comedies	Male and Female	English	Pitch contour; Blank-face	There is no specific ironical intonation. Pitch is used contrastively to signal irony or sarcasm. "Blank face" is a significant visual marker of irony or sarcasm.
Bryant and Fox-Tree (2005)	Ironic and nonironic	Dry, dripping, and nonironic utterances from Bryant and Fox-Tree (2002)	Male and Female	English	Pitch (F0 mean, range, variability); Amplitude variability; Speech rate	There is no specific "ironic tone of voice." Listeners interpret verbal irony by combining various cues, including contextual information, rather than relying on a distinct set of vocal cues.
Tepperman et al. (2006)	Sarcasm vs Sincere	Posed expressions filtered from Switchboard and Fisher corpora ("yeah right")	Male and Female	English	Contextual features (laughter, pauses); prosody (pitch, energy, duration); spectral features (MFCCs)	Prosodic features alone are insufficient for detecting sarcasm, and combining prosodic, spectral, and contextual features provides the best results.
Cheang and Pell (2007)	Sarcasm, humor, sincerity, neutrality	Posed expressions	Male and Female	English	Pitch (F0 mean, standard deviation, range); Amplitude (mean, range); speech rate; Harmonics-to-Noise Ratio; One-third octave spectral values	Sarcasm characterized by reductions in mean F0, HNR, and F0 standard deviation; changes in resonance and speech rate in certain contexts.
Regel (2009)	Ironic and nonironic	Discourse context + Target sentences	Female	German	Time (Duration: total sentence, sentence beginning, sentence ending); Pitch (F0 onset, F0 minimum, F0 maximum, F0 offset); Amplitude (Overall intensity, starting intensity $\geq$ 60 dB)	Ironic prosody differed from normal prosody: longer sentence duration, higher pitch at onset, lower pitch at offset, and consistently lower intensity. Context and prosody facilitated irony detection.

Author & Year	Attitude	Database	Gender	Language	Features	Result
Cheang and Pell (2009)	Sarcasm, humor, sincerity, neutrality	Keyphrases, sentences, combined sentences; Posed expressions	Male and Female	Cantonese & English	Pitch (F0 mean, range); Amplitude (mean, range); speech rate; Harmonics-to-Noise Ratio	Sarcastic utterances in Cantonese are produced with an elevated mean fundamental frequency (F0) and reductions in amplitude and F0 range, distinguishing them from sincere utterances. Sarcasm was spoken with a slower speech rate and a higher Harmonics-to-Noise Ratio (HNR). In contrast, English speakers lowered the mean F0 to mark sarcasm.
Scharrer et al. (2011)	Ironic criticism	Acted	Female	German	F0; vowel duration; intensity	Ironic criticism is marked by lower F0, increased vowel duration, and higher energy levels.
Rakov & Rosenberg (2013)	Sarcasm vs Sincere	Spontaneous speech recordings from Daria Sarcasm Corpus	Male and Female	English	Pitch (mean, range, standard deviation); intensity (mean, range); speaking rate; prosodic contours	Sarcastic speech is characterized by reduced pitch range, shallow intensity changes, and distinct prosodic contours. Achieved 81.57% accuracy in sarcasm detection.
Rao (2013)	Sarcasm and sincerity	Posed sarcasm	Male and Female	Mexican Spanish	Sentence-level: Pitch (F0 mean and range), speech rate; word-level: stressed syllable duration, F0 movement, stressed vowel intensity	Sarcasm led to decreases in speech rate and F0 mean, and increased stressed syllable length in attitudinally relevant words.
Niebuhr (2014)	Sarcasm, Sincere	Posed sarcasm	Male and Female	German	Pitch (mean, min, max, range); Amplitude (mean, min, max, range); Harmonic difference (H1–H2); Total duration	Sarcastic irony is expressed by longer utterance durations, lower and flatter F0 contours, and a lower intensity level.

Author & Year	Attitude	Database	Gender	Language	Features	Result
González-Fuente et al. (2016)	Irony literal	Story frameworks with target sentences and last-word analysis (Posed Irony)	Female	French	Pitch (mean, variability); Intensity (mean); intonation pattern (tonal nuclear configuration)	Speakers tended to interpret utterances as ironic when all acoustic features were presented together. Word lengthening and specific intonation patterns play a more significant role in conveying irony than pitch.
Braun & Schmiedel (2018)	Sarcasm / Kind irony / Sincere	Short scripted scenarios with keywords (Posed expression)	Male and Female	German	Pitch (mean, range, SD); intensity; duration; speech rate	Sarcasm: longer duration, lower pitch, less melodious; Kind irony: longer duration, higher pitch, more melodious. Listeners recognized irony above chance (72%).
Burkhardt et al. (2018)	Irony, anger	Spontaneous speech recordings by Sentiment Analysis / Acoustic Emotion Classification discrepancy	Male and Female	German	ComParE and eGeMAPS acoustic feature sets	Irony detected with 69.3% UAR. Prosodic features, especially pitch and intensity, contributed effectively.
Chen & Boves (2018)	Sarcasm, Sincere	Spontaneous speech recordings by simulated telephone conversation task	Male and Female	British English	Pitch (mean, max, min, span); duration; pitch contours; Functional Principal Components (FPCs)	Sarcasm is marked by longer word duration and flatter pitch contours, especially in key words. Female speakers used lower mean pitch, while males showed more variation in duration. Prosodic features predicted sarcasm with 70.4% accuracy.
Castro et al. (2019)	Sarcasm / non-sarcastic	Spontaneous multimodal recordings from TV (MUStARD)	Male and Female	English	BERT-based sentence representation; MFCC; Mel spectrogram; spectral centroid; visual features from video frames using a pre-trained ResNet-152 model	Multimodal information significantly enhances sarcasm detection performance; multimodal cues reduced relative error rate by up to 12.9% in F-score compared to individual modalities.

Author & Year	Attitude	Database	Gender	Language	Features	Result
Chauhan et al. (2020)	Sarcasm, sentiment, emotion	Multi-modal conversational dataset (MUStARD) (+3 sentiments + 9 emotions)	Male and Female	English	Text features (FastText embeddings, BiGRU, attention); visual features (average frames, speaker info); acoustic features (average frames, speaker info)	Incorporating both sentiment and emotion information improved the overall accuracy and effectiveness of sarcasm detection.
Bedi et al. (2021)	Sarcasm detection and humor classification	Hindi-English code-mixed dataset (MaSaC)	Male and Female	Romanized Hindi with English words	Pre-trained FastText multilingual embeddings; utterance-level hierarchical attention; MFCC	Hierarchical and contextual attention mechanisms improved performance, highlighting the importance of leveraging both textual and acoustic features.
Aguert (2021)	Irony / non-irony	Posed sarcasm key phrase	Male and Female	French	Time (duration, pauses); Pitch (mean, variability); Intensity (mean); Speech rate; facial cues (eyebrows, gaze, expressive mouth movements)	Ironic speakers produced more pauses and different vocalizations, had a slower speech rate, and spoke with lower intensity. Facial cues (especially eyebrow flashes and expressive mouth movements) were more effective than vocal cues.
Ray et al. (2022)	Sarcasm, sentiment, emotion	MUStARD++ (Valence/arousal; Propositional/Embedded/Like-prefixed/Illocutionary)	Male and Female	English	Textual features (BART-large); acoustic features (MFCC, Mel spectrogram); visual features (ResNet-152 from key frames)	Multimodal fusion achieved 72% F1 for sarcasm detection; emotion recognition revealed distinct valence and arousal patterns in sarcastic utterances.
Rao et al. (2022)	Sarcasm versus sincerity	Posed sarcasm; Discourse context + Target sentences	Male and Female	English-dominant English-Spanish bilinguals	Pitch (mean, range); Speech rate	Sarcasm was produced with slower speech rate and lower pitch mean. Spanish had higher f0 mean/range than English, with variation by age and dominance.

Author & Year	Attitude	Database	Gender	Language	Features	Result
Bolyanatz et al. (2023)	Jocularity, rhetorical questions, hyperbole, understatements, and sarcasm	Spontaneous speech recordings (semi-structured interview)	Female	Chilean Spanish	Pitch (mean, range for F0, F1, F2); Time (duration); H1*–H2*; HNR	Jocularity, hyperbole, and rhetorical questions significantly differed from baseline utterances. Jocularity/hyperbole showed wider pitch range; rhetorical questions exhibited higher overall pitch; all three showed lower HNR.
Li et al. (2024)	Sarcasm	MUStARD++ (utterance + key phrases annotation)	Male and Female	English	Pitch (mean, range); amplitude (mean, range); duration	Sarcastic utterances with salient semantic cues relied less on prosodic modulation, while those with fewer semantic indicators emphasized prosodic features (e.g., lower mean F0 and higher mean amplitude).
Caucci et al. (2024)	Sarcastic tone	Posed sarcasm	Male and Female	English	Pitch (mean, variability); Amplitude (mean, variability); Speech rate	Sarcastic speech had slower rates, lower amplitude, less amplitude variability, and more pitch variability.

2.5 Affective Algorithmic Composition

Affective Algorithmic Composition (AAC) constitutes a research domain concerned with the computational governance of emotional expression in automatically generated music, with the overarching objective of eliciting or systematically modulating specific affective responses in listeners. As a specialization within the broader field of automated music generation, AAC investigates the mechanisms by which musical structure, performance parameters, and compositional processes may be systematically manipulated to convey emotional meaning with precision and consistency (Williams et al. 2015; Juslin 2013).

Affective music generation systems are usually classifies into four major paradigms: rule-based systems, evolutionary optimization approaches, data-driven neural models, and conditional neuro-audio generation architectures. This taxonomy is broadly consonant with recent comprehensive surveys of AI-based affective music generation, which draw a principled distinction between *knowledge-driven* and *learning-based* strategies for affective control (Dash and Agres 2024). Rule-based systems instantiate explicit, declarative mappings between musical surface features and emotional dimensions, typically grounded in established music psychology and music theory. Evolutionary approaches reformulate affective alignment as a combinatorial optimization problem, iteratively refining musical material through fitness-based selection against predefined affective targets. Data-driven neural and deep learning models, by contrast, acquire implicit, distributed representations of the relationship between musical structure and affective labels from large-scale corpora, affording substantially greater stylistic diversity and generative flexibility than their knowledge-engineered counterparts. Extending this trajectory, conditional neuro-audio generation systems learn joint latent representations of natural language and acoustic content, thereby enabling affective intent to be specified through open-ended linguistic prompts rather than constraining it to the fixed geometry of predefined emotion taxonomies (Forero, Bernardes, and Mendes 2023).

2.5.1 Rule-Based Methods

In automated music generation, rule-based methods constitute some of the earliest and most interpretable approaches to affective control. These systems encode explicit relationships between musical structure and emotional intent through predefined compositional rules, which may take the form of mathematical formulations, logical constraints, or heuristic procedures (e.g., conditional if–then–else mappings). Within an automated music generation framework, such rules govern the selection, transformation, and combination of musical parameters in order to convey specific affective qualities.

Early works in this area explored real-time music generation for interactive and virtual environments, in which high-level affective descriptors were mapped onto lower-level musical parameters. Robertson et al. (1998) proposed a real-time generative system in which user input, expressed as a continuous *tension* parameter, governed multiple layers of musical organization simultaneously. The system employed a modular architecture encompassing high-level form generation, rhythmic

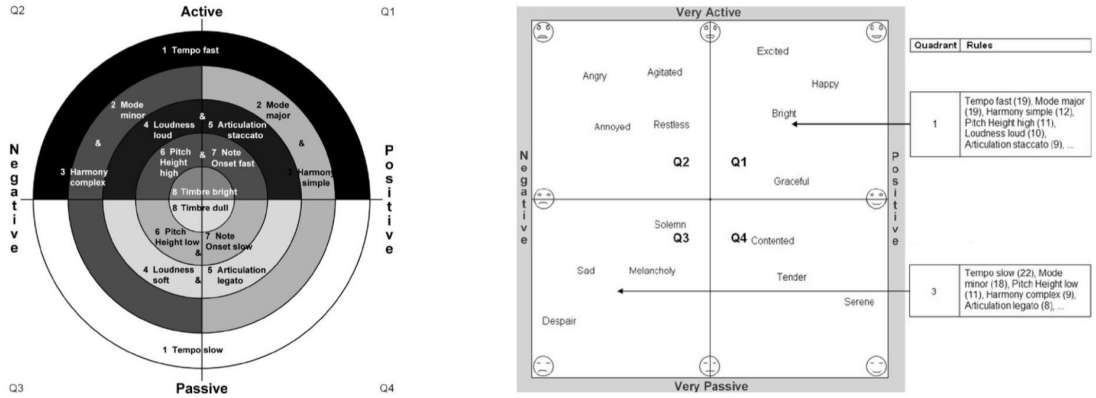


Figure 2.3: Rule-based mapping between musical parameters and emotional expression in the valence–arousal space. The diagram illustrates how compositional features—including tempo, mode, harmony, pitch register, articulation, and timbre—are associated with discrete affective quadrants (e.g., active–positive, passive–negative), enabling systematic transformation of musical material to induce target emotional states. Adapted from Livingstone et al. (2010).

and dynamic control, melody and harmony generation, and a dedicated playback component. By projecting the tension parameter onto internal musical variables, the system dynamically modulated the perceived intensity and affective character of the generated output during live interaction.

Subsequent research sought to formalize and theorize the relationship between musical surface structure and affective dimensions. Wallis (2008) introduced a rule-based affective music generation system grounded in music-theoretical principles derived from empirical studies of musical expression, most notably those of Gabrielsson and Lindström (2001). Such approaches are typically operationalized through explicit parametric mappings between compositional features and positions in the valence–arousal space, as illustrated in Figure 2.3.

Rule-based affective control has been extended to cross-modal creative systems that synthesize musical and linguistic meaning. *TransProse*, developed by Davis and Mohammad (2014), generates piano melodies from literary texts by coupling lexicon-based sentiment analysis with a set of predefined compositional rules. The system partitions a novel into discrete textual segments, assigns an emotional label to each unit, and instantiates corresponding melodic material by controlling parameters such as scale, tempo, octave range, and pitch selection. Positive textual segments are rendered in major mode, whereas negative segments are rendered in minor mode. The system was evaluated through qualitative analysis of musical outputs generated from nine canonical novels.

Probabilistic architectures have also been integrated with rule-based affective control to support dynamic musical adaptation. Williams et al. (2015) proposed an affectively driven music generation system for adaptive game soundtracks, grounded in a second-order Markov model. In this framework, affective trajectories defined along the arousal and valence dimensions of the circumplex model of affect are projected onto musical parameters including rhythmic density, tempo, modality, articulation, mean pitch range, and spectral range. A learned transition matrix enables

smooth, real-time adaptation of the generated music in response to evolving affective targets.

More architecturally complex frameworks integrate rule-based affective reasoning with evolutionary and stochastic generative components. Scirea et al. (2017) introduced *MetaCompose*, a real-time music generation framework for interactive video game environments, which unifies stochastic chord generation, evolutionary melody optimization, and rule-based affective transformation within a single pipeline. A dedicated expert system, termed the Real-Time Affective Music Composer, modifies musical attributes including dynamic level, timbre, rhythmic density, and harmonic dissonance to align generated output with target affective states defined within Russell's circumplex model of affect.

2.5.2 Evolutionary Algorithm-Based Systems

Evolutionary algorithms (EAs) constitute a class of stochastic optimization techniques inspired by Darwinian principles of natural selection and genetic variation, and have been widely adopted in affective music generation to iteratively refine musical material toward desired emotional or stylistic targets. In these systems, candidate musical solutions are encoded as chromosomes, whose constituent genes may represent individual notes, chords, rhythmic patterns, bars, or larger-scale structural units. Through successive generations, mutation and crossover operators introduce controlled variation into the population, while a fitness function evaluates each candidate according to affective, stylistic, or perceptual criteria, determining which individuals are selected for further evolution.

The earliest systems in this paradigm relied on explicit user judgment as the primary fitness signal. Kim and André (2004) proposed an interactive evolutionary music composition system that integrates a perceptual user interface with genetic optimization. Users explicitly evaluated generated rhythms using affective descriptors—*relaxing*, *neutral*, or *disquieting*—allowing the system to construct a model of individual preference. Once this preference model was established, physiological signals were substituted as an implicit fitness measure, enabling affect-responsive refinement of the generated material without requiring continuous conscious evaluation. Building on the interactive paradigm, Zhu, Wang, and Wang (2008) introduced an Interactive Genetic Algorithm (IGA) for emotional music generation grounded in the KTH rule system, a formal computational model of expressive performance principles in Western classical, jazz, and popular music, which governs parameters such as phrasing, articulation, and tonal tension through calibrated weighted coefficients. By evolving these performance parameters interactively, the system generated musical performances aligned with target emotional states while preserving stylistically coherent expressive nuance.

The scope of evolutionary optimization was subsequently extended from performance expression to harmonic structure. Xu, Wang, and Li (2010) proposed an emotional harmony generation system combining predefined harmonic rules with an interactive genetic algorithm. Their framework distinguishes analytically between *harmony composition rules*, which govern chordal

structure and voice-leading, and *harmony emotion rules*, which establish mappings between harmonic features and discrete emotional categories such as happiness and sadness. Fitness evaluation incorporated direct user judgment via IGA, and both subjective listening tests and objective acoustic analyses confirmed the system's capacity to generate harmonically and affectively distinguishable progressions. Extending the interactive EA paradigm to distributed multi-user settings, Nomura and Fukumoto (2018) introduced a Distributed Interactive Genetic Algorithm (DIGA) for generating short piano melodies with controllable perceptual attributes, operationalized as perceived *brightness*. Independent melodic populations were evolved in parallel by multiple users, with fitness assessed on a seven-point Likert scale anchored at *extremely dark* and *extremely bright*. Experimental results demonstrated improved fitness convergence and increased inter-individual melodic consistency when cross-user population exchange was enabled, suggesting that distributed co-evolution can partially mitigate the evaluation fatigue and subjective inconsistency that characterize single-user IGA designs.

More recently, the fitness signal has been decoupled from real-time user evaluation through the use of statistical mood templates. Rocha de Azevedo Santos, Silla Jr., and Costa-Abreu (2021) proposed a genetic algorithm-based methodology for procedural piano music composition in which MIDI templates representing contrasting affective profiles—including *happy* and *sad*—served as the optimization target. Randomly initialized musical populations were evolved until their aggregate statistical features converged with those of the target templates. Human listener evaluations confirmed that the generated pieces conveyed the intended affective profiles; however, they were consistently perceived as computer-generated, with raters identifying rhythmic regularity and insufficient expressive nuance as the primary markers of artificiality. This finding delineates a persistent boundary condition for EA-based affective generation: statistical convergence with affective templates does not, in itself, guarantee perceptual authenticity, motivating the turn toward data-driven neural approaches capable of capturing higher-order stylistic dependencies.

2.5.3 Neural networks

Neural networks, especially Long Short-Term Memory networks, have been pivotal in AAC for generating emotion-driven music. Systems such as SentiMozart (Madhok, Goel, and Garg 2018) use a two-model approach, where the first model classifies facial expressions into emotion categories and the second generates music based on this information. Ferreira and Whitehead (2019) used a multiplicative LSTM model for generating music with specific emotions, while Zhao et al. (2019) introduced a conditional Biaxial LSTM model for generating piano pieces with controllable emotions. Ferreira, Lelis, and Whitehead (2020) introduced Bardo Composer, a system that generates music for tabletop role-playing games based on players' moods. K. Zheng et al. (2022) presented EmotionBox, a music-element-driven generator trained using a gated recurrent unit (GRU) to produce music with emotions like happy, sad, and peaceful.

Variational Autoencoders (VAE) are another widely used neural network architecture in AAC. Google's MusicVAE (Roberts et al. 2018) employs a hierarchical VAE to learn long-term musical structure, enabling coherent multi-bar generation. Building on this approach, Tan and Herremans

(2020) proposed Music FaderNets, which use Gaussian Mixture VAEs to manipulate low-level features like rhythm and note density to generate emotionally controlled music. Generative Adversarial Networks (GANs) are used in AAC to generate emotionally expressive music. Huang et al. (2022) proposed ECMG, a GAN-based melody generation model incorporating emotion constraints, producing high-quality melodies with specific emotions. The CVAE-GAN model (Huang and Huang 2020) also utilizes emotion tags to generate music segments corresponding to specific emotions.

2.5.4 Conditional Neuro-Audio Generation

Recent advances in large-scale generative modeling have fundamentally reoriented affective music generation by enabling systems to synthesize audio directly from natural language prompts. Unlike the rule-based and evolutionary approaches surveyed in Sections 2.5.1 and 2.5.2, which depend on explicit, expert-defined mappings between emotional labels and discrete musical parameters, conditional neuro-audio generation models learn implicit, high-dimensional relationships between linguistic descriptions and acoustic structure from large-scale multimodal corpora.

Early large-scale systems demonstrated that stylistic and affective qualities could emerge from loosely specified textual conditioning, even in the absence of explicit emotion supervision. OpenAI's *Jukebox* (Dhariwal et al. 2020) exemplified this capacity by combining vector-quantized variational autoencoders (VQ-VAEs), which compress raw audio waveforms into discrete token sequences via a learned code-book, with autoregressive transformer models that predict subsequent tokens conditioned on high-level metadata, including genre, artist style, and lyrical content. Although not explicitly trained on affective objectives, *Jukebox* demonstrated that emotional and stylistic characteristics can be implicitly encoded within learned musical representations, emerging as a coherent byproduct of large-scale generative modelling rather than as a deliberately engineered property.

Subsequent architectures advanced this trajectory by strengthening the semantic alignment between textual input and generated audio output. Google's *MusicLM* (Agostinelli et al. 2023) introduced a hierarchical sequence-to-sequence architecture, integrating neural audio tokenization via SoundStream, self-supervised acoustic representation learning via w2v-BERT, and contrastive audio–text embedding alignment via MuLan, trained on large-scale paired music, text data. This multimodal representational hierarchy enables the generation of high-fidelity musical audio that adheres closely to detailed prompt specifications. Critically, *MusicLM* demonstrated a substantially improved capacity to translate affective language, including references to mood, atmosphere, and emotional intensity, into perceptually congruent musical structures. Evaluation combining human perceptual judgments with embedding-based metrics confirmed robust semantic and affective fidelity across a wide range of prompt types (see Figure 2.4).

Meta AI's *MusicGen* (Copet, Défossez, Gat, Nguyen, et al. 2023), developed within the *AudioCraft* framework (Copet, Défossez, Gat, Huang, et al. 2023), further consolidated text-to-music generation by employing a single-stage autoregressive transformer that predicts discrete audio tokens conditioned on text encoder representations, dispensing with the multi-stage hierarchical

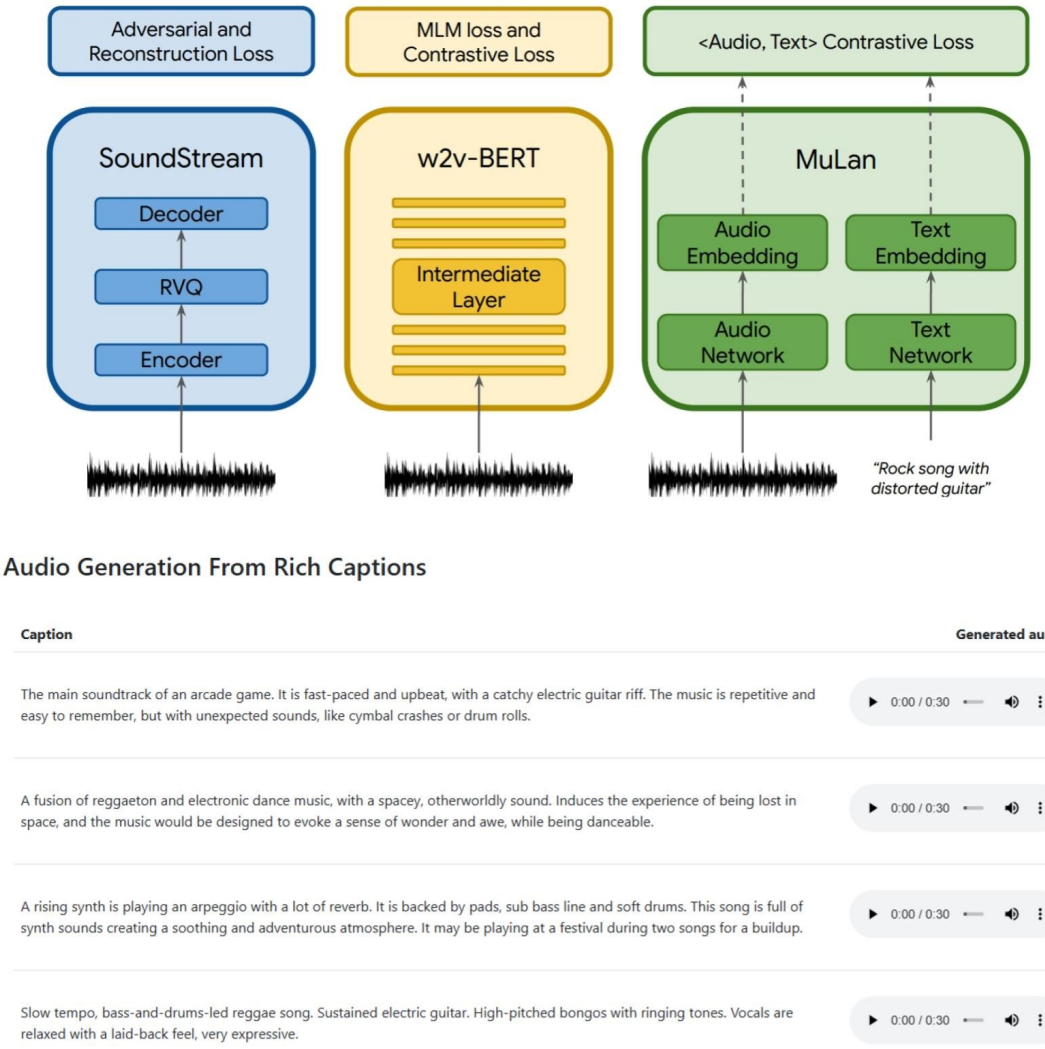


Figure 2.4: Overview of semantic music generation in *MusicLM*. The architecture integrates neural audio tokenization (SoundStream), self-supervised acoustic representation learning (w2v-BERT), and joint audio–text embedding via contrastive learning (MuLan) to align natural language descriptions with learned musical structure. This multimodal representational hierarchy enables coherent high-fidelity audio synthesis directly from rich natural language prompts, as illustrated by the mapping of descriptive captions to synthesized musical outputs. Adapted from Agostinelli et al. (2023).

pipelines characteristic of earlier systems. Despite this architectural parsimony, *MusicGen* demonstrated effective sensitivity to affective descriptors such as *melancholic*, *uplifting*, and *tense*, producing musically coherent outputs that reflect the intended emotional register. Its computational efficiency and prompt-level controllability render it particularly well-suited to interactive and real-time creative applications in which low-latency affective responsiveness is a primary constraint.

In parallel, diffusion-based architectures have emerged as a principled alternative strategy for conditional neuro-audio generation, replacing the autoregressive token prediction objective with an iterative denoising process grounded in score matching. *Noise2Music* (Huang et al. 2023) employs a cascaded diffusion process in which a sequence of progressively higher-fidelity diffusion models transforms Gaussian noise into structured audio conditioned on textual prompts, leveraging large-scale pseudo-captioning pipelines that associate musical excerpts with affective natural language descriptions. *Riffusion* (Forsgren and Martiros 2022) adapted text-to-image latent diffusion principles to the audio domain by generating mel-spectrograms from textual prompts and converting them to waveforms via a vocoder, enabling continuous interpolation between affective and timbral conditions in the spectrogram latent space. Commercial systems such as Stability AI's *Stable Audio* (Evans et al. 2024; Stability AI 2023) integrate latent diffusion models with explicit timing and duration conditioning, facilitating the generation of emotionally aligned audio suitable for synchronization with narrative or visual media.

Across all of these architectures, affective control is mediated through the embedding of natural language in high-dimensional latent spaces learned from extensive multimodal corpora. Evaluation frameworks combining human perceptual judgments with multimodal embedding similarity metrics, including Contrastive Language–Audio Pretraining (CLAP) scores (Elizalde et al. 2022), indicate strong alignment between generated audio and intended affective qualities, while also revealing the interpretive opacity that distinguishes emergent latent representations from the transparent parameter-to-emotion mappings of their rule-based predecessors.

2.6 Affective Virtual Environment Generation

Affective Virtual Environments (AVEs) are immersive systems designed to elicit or modulate emotional states through the deliberate configuration of perceptual variables. Virtual environments have been shown to elicit measurable emotional responses under controlled conditions (Slater and Sanchez-Vives 2016; Felnhofer et al. 2015; Chirico et al. 2017), and most studies employ a combination of physiological assessments (e.g., EEG, ECG, galvanic skin response, heart rate variability) and standardized self-report instruments (e.g., SAM, Likert-based valence–arousal scales) to validate emotional induction (Somarathna, Bednarz, and Wang 2022). From a computational standpoint, AVEs may be implemented through distinct representational and generative strategies, including fully modeled three-dimensional scenes, image-based immersive projections (e.g., panoramic or 360° environments), and learned data-driven pipelines.

A distinctive dimension of the present work, however, concerns not affective congruence but affective *divergence*: environments in which the acoustic and visual channels intentionally convey opposed emotional valences in analogy to prosodic irony described in 2.4. The theoretical grounding for this strategy derives from contrapuntal sound theory, as advanced by Eisenstein, Pudovkin, and Alexandrov (1985) within the context of Soviet montage. Eisenstein conceptualized sound as a medium capable of deliberate *non-synchronization* with the image, thereby generating emergent meanings irreducible to either channel in isolation. This vocabulary was significantly refined by Chion (1994), whose taxonomy of *empathetic* and *anempathetic* sound formalizes the affective poles of the sound-image relationship. While empathetic music tracks the emotional contour of a scene, anempathetic sound proceeds with a conspicuous indifference to the narrative action, a detachment that, Chion argues, heightens the sense of the tragic through emotional displacement.

The deployment of such mismatch for ironic effect has a rich history in cinema. Gorbman (1987) distinguishes between *blunt* irony, such as Buñuel's scoring of a parodic Last Supper with Handel's *Hallelujah Chorus* in *Viridiana* (1961), and *structural* irony, as seen in Kubrick's *2001: A Space Odyssey* (1968), where the juxtaposition of waltzing spacecraft with Strauss generates irony through the clash of cosmic vastness and ballroom grace. In dark comedy, from *Fargo* (1996) to *The Death of Stalin* (2017), incongruent music serves as a multi-functional device to manage affective distance and signal a black-comedic register (Ireland 2018). Ireland (2017) further notes that when such ironic readings become stylistic expectations, the anempathetic music paradoxically becomes congruent with the director's established aesthetic code. This cross-modal irony is not merely a cinematic convention but a constitutive perceptual force in immersive environments: McArthur (2016) demonstrates that even subtle mismatched audiovisual cues can disrupt immersion in VR, while Malpica et al. (2023) establish that thematic dissonance lowers perceived atmosphere without necessarily impairing self-reported immersion.

2.6.1 Rule-Based Methods

Fully modeled three-dimensional environments constitute the most classical and structurally explicit paradigm for AVE generation. In this approach, spatial geometry, illumination, material

properties, object placement, narrative structure, and interaction logic are deliberately authored according to predefined mappings between perceptual variables and targeted affective states. Emotional modulation is therefore implemented through deterministic design rules grounded in psychological theory rather than through statistical learning.

An early example is the VR Mood Induction Procedure developed by Baños et al. (2006), in which a virtual park environment was systematically modified through controlled adjustments of lighting conditions, color tone, ambient sound, and environmental dynamics to induce sadness, happiness, anxiety, and relaxation. Emotional states were validated using self-report and psychophysiological measures, illustrating explicit stimulus–affect mapping within a fully modeled 3D setting. Subsequent work expanded this rule-based paradigm toward discrete emotional elicitation in immersive contexts. Felnhofer et al. (2015) designed multiple virtual park scenarios to evoke joy, sadness, anger, anxiety, and boredom. Emotional differentiation was achieved through structured manipulation of weather conditions, ambient soundscapes, lighting intensity, and environmental content, demonstrating how compositional environmental variables can be strategically configured to induce targeted affective responses.

Moving toward dimensional models of emotion, Marín-Morales et al. (2018) constructed four architecturally distinct immersive virtual rooms by systematically manipulating illumination, chromatic composition, and geometric enclosure to correspond to specific quadrants of the valence–arousal circumplex model. The affective validity of these environments was confirmed through physiological measurements and self-report instruments, reinforcing the capacity of controlled spatial parameters to bias emotional experience along continuous affective dimensions.

More recent methodological contributions have formalized rule-based affective design frameworks. Dozio et al. (2021) introduced a validated database of interactive affective virtual environments, explicitly constructed to differentiate discrete emotional categories. Building upon this foundation, Dozio et al. (2022) proposed a structured methodology for affective VR design grounded in the Valence–Arousal–Dominance model, authoring and experimentally validating ten targeted emotional environments.

2.6.2 Image-Based Representations

Image-based approaches generate immersive environments without fully reconstructing geometric structure. Instead of explicitly modeling three-dimensional geometry, these systems rely on captured visual data such as panoramic images, 360° video, cubemaps, or environment maps to produce immersive experiences (Debevec, Taylor, and Malik 1996). Because scene structure is not parametrically editable at the geometric level, affective modulation typically operates through global appearance manipulations, including color grading, luminance adjustment, filtering, atmospheric overlays, or audio-visual augmentation.

One of the earliest uses of immersive image-based stimuli for affective research involved VR Mood Induction Procedures based on panoramic and 360° environments. For instance, extensions of the VR-MIP paradigm employed immersive panoramic content to induce emotional states while

preserving photorealistic context. Emotional differentiation was achieved through manipulation of color tone, ambient audio, and scene composition rather than geometric restructuring.

More systematic image-based affective datasets have emerged in recent years. Li et al. (2021) introduced a 360° video database annotated along valence–arousal dimensions, enabling the study of affective responses to immersive visual stimuli without requiring interactive scene modeling. Similarly, Suhaimi, Mountstephens, and Teo (2020) constructed a 360° immersive video dataset including romantic, horror, skydiving, and nature scenarios, each evaluated along dimensional affect scales. These works demonstrate how emotional induction can be achieved through curated immersive recordings rather than synthetic modeling.

Recent multimodal affective VR studies, such as those reviewed by Somarathna, Bednarz, and Wang (2022), further illustrate the role of immersive video-based stimuli in emotion recognition research. In these paradigms, affective responses are measured via physiological signals (e.g., EEG, heart rate variability, galvanic skin response) while participants are exposed to 360° videos designed to evoke specific valence–arousal states. Although these environments lack geometric manipulability, they provide high ecological validity and strong perceptual realism.

Image-Based Rendering (IBR) techniques extend this approach by synthesizing novel viewpoints from image collections (Shum and Kang 2000). While computationally efficient and capable of photorealistic reconstruction, IBR systems provide limited direct control over spatial topology or object-level geometry. As a result, affective modulation in image-based systems primarily operates at the level of global visual appearance and scene semantics rather than structural transformation.

2.6.3 Procedural Encoding

Procedural systems generate environments algorithmically through rule sets rather than static authoring. Terrain formation, vegetation growth, architectural layouts, object density, lighting variability, and atmospheric conditions may be governed by parameterized generative rules. Instead of manually modeling a fixed scene, designers define constraints and transformation functions that produce spatial configurations dynamically. In affective contexts, such systems enable high-level emotional parameters (e.g., tension, serenity, unpredictability, vastness) to propagate across environmental structures in a scalable and systematic manner.

One early conceptual foundation for procedural affective generation can be traced to appraisal-based agent architectures. Reilly (1996) proposed rule-driven emotional agents within the Oz Project, where affective states emerged from computational appraisal processes derived from the OCC model. Although focused on character behavior, this framework established how emotional dynamics could be generated procedurally rather than pre-scripted. Similarly, Martinho and Paiva (2000) developed emotionally-driven virtual characters whose internal appraisal mechanisms influenced narrative flow and environmental interaction, demonstrating algorithmic emotion regulation in interactive spaces.

Procedural affective modulation is also evident in adaptive VR systems that regulate environmental parameters in response to user performance or physiological signals. For example, flow-oriented systems such as Magic XRoom (Mousavi, Vellido, and Liapis 2024) dynamically adjust task difficulty and environmental constraints according to predefined mappings between challenge and affective engagement. Here, emotional state is not embedded in a fixed scene but emerges from continuous procedural recalibration of environmental demands.

In experimental affective VR, frustration-based anger scenarios and threat-driven fear environments frequently rely on algorithmic manipulation of event frequency, obstacle distribution, or failure loops. The affective VR database introduced by Dozio et al. (2021) includes environments where emotional induction (e.g., anger or fear) is achieved through systematic procedural rules such as increasing task interruption, constraining user control, or modulating sensory unpredictability. Rather than altering geometry alone, these systems manipulate interaction logic and temporal structure to generate targeted affective states.

Procedural encodings are also common in VR games designed for emotion classification. For instance, Meuleman, Rudrauf, and Lamy (2018) employed interactive VR gameplay with dynamically evolving scenarios to elicit multi-componential emotional responses, while Granato, Gadia, and Marfia (2020) used racing-game environments where speed, risk density, and competitive dynamics were governed algorithmically to modulate arousal and valence. In such systems, emotional variability arises from procedural control of environmental tempo, risk exposure, and performance feedback.

2.6.4 Data-Driven Representations

Recent advances in neural scene representations introduce a data-driven alternative to explicit geometric modeling and procedural rule encoding. Instead of manually authoring meshes or defining algorithmic transformation rules, implicit neural representations encode geometry, appearance, and temporal dynamics as continuous functions parameterized by neural networks. Scene properties are learned from data distributions and stored in high-dimensional latent parameter spaces. In this paradigm, affective modulation may be achieved through latent-space conditioning, generative prompting, or optimization toward perceptual priors statistically associated with target affective states.

A foundational example of this shift is the Neural Radiance Field (NeRF) model proposed by Mildenhall et al. (2020). NeRF represents a scene as a continuous function mapping spatial coordinates and viewing direction to volumetric density and view-dependent radiance. Rather than storing explicit surface meshes or voxel grids, the scene is encoded as the parameters of a multilayer perceptron trained via differentiable volume rendering. Geometry emerges implicitly through density gradients optimized from multi-view images. Subsequent work, such as Instant Neural Graphics Primitives (Müller et al. 2022), introduced multiresolution hash encodings that dramatically accelerate training and rendering, making neural fields computationally viable for interactive systems. In affective virtual environments (AVEs), such representations permit

global appearance modulation—e.g., luminance distribution, spatial enclosure, or atmospheric density—through parameter optimization rather than geometric restructuring.

Complementing radiance-based representations, neural implicit surface methods reconceptualize geometry itself as a learned continuous function. Park et al. (2019) introduced DeepSDF, which models objects as signed distance functions parameterized by neural networks. Object surfaces are defined implicitly as zero-level sets of continuous fields, enabling compact storage and smooth interpolation across shape manifolds. Similarly, Occupancy Networks (Mescheder et al. 2019) learn probabilistic occupancy functions over continuous space, allowing topology to emerge from data rather than explicit mesh construction. Within AVEs, such latent geometric representations allow controlled interpolation between spatial configurations (e.g., open versus enclosed architectures or curved versus angular forms) that have been empirically associated with affective appraisal.

Temporal extensions of neural implicit models further expand their affective potential. D-NeRF (Pumarola et al. 2021) incorporates time as an additional input dimension and employs a learned deformation network that maps canonical spatial coordinates to time-conditioned configurations. Motion is therefore modeled through continuous deformation fields rather than external animation layers. This encoding of structure and dynamics within a unified differentiable framework aligns with affective modulation of temporal parameters such as pacing, oscillation, or environmental rhythm. Rather than scripting motion intensity to regulate arousal, dynamic patterns may be embedded directly within learned spatiotemporal representations.

Generative neural scene representations further decouple environment creation from direct observation. DreamFusion (Poole et al. 2023) demonstrated that coherent three-dimensional objects can be synthesized by distilling priors from large-scale 2D diffusion models into neural radiance fields using Score Distillation Sampling (SDS). Rather than requiring paired text–3D datasets, DreamFusion leverages a pretrained 2D text-to-image diffusion model as a supervisory signal to optimize a 3D representation such that its rendered views align with textual prompts. This approach bypassed the scarcity of large-scale text–3D data and established a viable pathway from semantic prompts to volumetric scene representations.

Subsequent refinements improved resolution, geometric fidelity, and computational efficiency. Magic3D (Lin et al. 2023) introduced a two-stage coarse-to-fine pipeline, combining low-resolution NeRF optimization with higher-resolution mesh refinement guided by diffusion priors. Zero-1-to-3 (Liu, Zhou, Efros, et al. 2023) further demonstrated that single-view image conditioning can generate consistent multi-view renderings, advancing spatial coherence from minimal visual input. While these systems represent a substantial advance in semantic-to-spatial generation, their outputs typically consist of static 3D assets or pre-rendered scenes. They lack persistent dynamics, real-time interactivity, and the temporal continuity required for fully manipulable environments.

Concurrently, video generation research addressed the temporal dimension of generative modeling. Imagen Video (Ho et al. 2022) extended diffusion-based architectures to high-definition text-to-video synthesis through cascaded spatiotemporal diffusion models. Make-A-

Video (Singer, Polyak, Hayes, et al. 2022) leveraged pretrained text-to-image models alongside unlabeled video data to learn motion priors without requiring text–video pairs. AnimateDiff (Guo et al. 2023) introduced modular motion adapters that enable controllable animation generation from pretrained text-to-image diffusion systems. Although these models produce temporally coherent clips and implicitly learn motion regularities and physical plausibility from large-scale video corpora, their outputs remain non-interactive sequences. The generated video constitutes a predetermined stream of frames rather than a simulation responsive to user input.

The integration of linguistic control (text-to-image), spatial modeling (text- or image-to-3D), and temporal synthesis (text-to-video) has recently converged toward a new paradigm: interactive world generation. Unlike static asset creation or short video clips, this emerging class of systems aims to generate persistent environments with latent physical structure capable of responding to user actions in real time.

A seminal contribution in this direction is Genie (Bruce et al. 2024), a foundation world model developed by DeepMind. Genie is trained on large-scale internet video data and learns a compressed, action-conditioned latent space capable of modeling appearance, motion, and environmental dynamics. From a single input image, the system can generate an interactive environment in which user actions (e.g., movement commands) produce consistent and temporally coherent responses. Rather than relying on explicitly programmed physics engines, Genie demonstrates that interactive dynamics can be learned directly from observational video corpora. This represents a conceptual shift from content generation toward generative simulation.

Building upon similar principles, Tencent's *Hunyuan World* (Tencent Hunyuan Team 2024) introduces an open-source framework for interactive 3D world generation that moves beyond static scene synthesis toward persistent, navigable environments. The model is trained on large-scale video datasets, allowing it to learn not only visual appearance but also implicit spatial structure and temporal consistency. Hunyuan World encodes video data into a latent representation that captures correlations between consecutive frames, enabling the model to infer depth, geometry, and motion dynamics without relying on explicit 3D supervision. As shown in image 2.5, given a text prompt, the system first generates a panoramic or global scene representation that establishes semantic layout and environmental context. This representation is then used to synthesize consistent multi-view renderings, ensuring that objects, lighting, and spatial relationships remain stable across viewpoints.

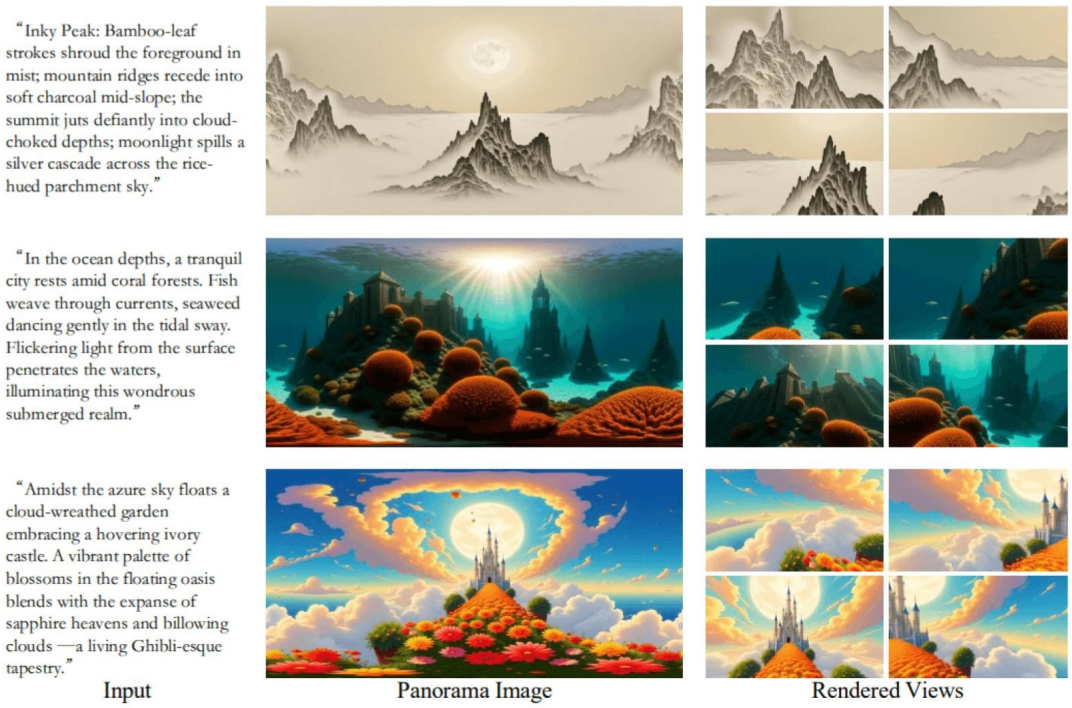


Figure 2.5: Semantic-to-spatial generation pipeline in Hunyuan World. A natural language prompt (left) is first mapped into a latent semantic representation, which is used to synthesize a globally coherent panoramic scene (center). From this representation, the model generates consistent multi-view renderings (right), preserving object geometry, lighting, and spatial relationships across viewpoints. By learning spatiotemporal correlations from large-scale video data, the system enables persistent, navigable environments that approximate interactive world simulation from textual input.

Chapter 3

Methodology

This study operates at the intersection of media art and computer science, employing a mixed-methods approach that bridges experimental computational modeling with practice-based artistic inquiry. The research is structured as a recursive dialogue between these two modes: the experimental design governs the systematic development and evaluation of bimodal speech emotion recognition models, while the practice-based methodology (Candy 2006) positions creative production and critical reflection as the primary sites for generating situated knowledge. In this framework, the computational analysis of verbal irony provides the structural foundation for the development of Affective Virtual Environments (AVE), while the artistic applications function as both experimental case studies and speculative probes that test the boundaries of these technological frameworks.

The research follows an iterative procedure to develop an end-to-end system in which *The Ironic Machine[s]* operates as speculative instances of an affective virtual environment. The AVE system is composed of two primary and functionally independent components: an **automatic emotion recognition module** and a **virtual environment generation module**, as illustrated in Figure 3.1. Through this AVE system, we investigate how multimodal speech emotion datasets can

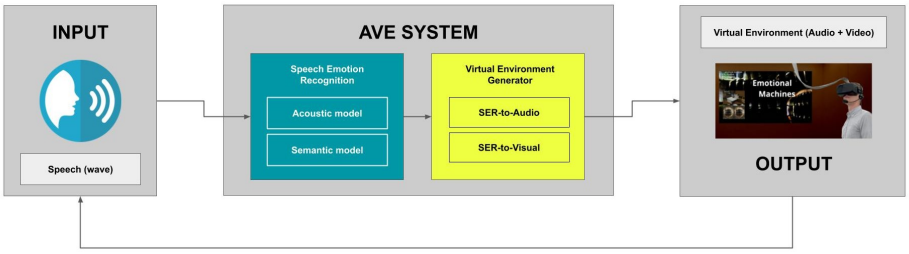


Figure 3.1: Architecture of the Affective Virtual Environment (AVE) System. The workflow illustrates the transformation of input speech into an immersive virtual environment. The process involves two primary mechanisms: an emotion recognition system, using acoustic and semantic models to classify affect, and a virtual environment generation system that maps these classifications to corresponding audiovisual attributes for VR rendering.

support the computational analysis of verbal irony. Furthermore, we define and evaluate a strategy for mapping ironic speech, understood as a divergence between semantic and prosodic modalities, into affective virtual environments. This approach integrates statistical inference with practice-based creative inquiry, giving rise to nine creative projects that function both as experimental case studies and as situated artistic proposals.

For transparency and reproducibility, the complete implementation of the experimental pipelines, together with the resulting artistic outputs, is publicly available in an open Git repository[1]. The project can also be explored through its accompanying web interface[2].

3.1 Emotion Recognition

For emotion recognition, we propose using speech as the leading interaction modality, which encodes meaning through both textual content and prosodic acoustic attributes.[3] The system operates through two independent machine learning pipelines that run concurrently to predict emotions from the **semantic** and **acoustic** channels. Each pipeline is implemented through a set of supervised classifiers that infer affective categories from either lexical content or prosodic features.

The classifiers evaluated in this study are organized as follows:

- **Semantic modality (text-based sentiment analysis):**
  - **Rocchio classifier:** a prototype-based method grounded in vector-space representations.
  - **LSTM classifier:** a recurrent neural network capable of modeling sequential linguistic dependencies.
  - **BERT classifier:** a transformer-based model leveraging contextualized language representations.
- **Acoustic modality (speech emotion recognition):**
  - **MLP classifier:** a feed-forward neural network operating on fixed-length acoustic descriptors.
  - **CNN classifier:** a convolutional architecture designed to capture local patterns in the feature space.
  - **CNN+LSTM classifier:** a hybrid model combining convolutional feature extraction with temporal sequence modeling.

1. https://github.com/jfforero/The-Ironic-Machines
2. https://jfforero.github.io/The-Ironic-Machines/
3. While many spoken languages can be transcribed into textual representations, not all communicative systems possess a standardized or explicit semantic encoding in written form. However, all spoken languages convey meaning through a combination of symbolic or contextual interpretation and acoustic features.

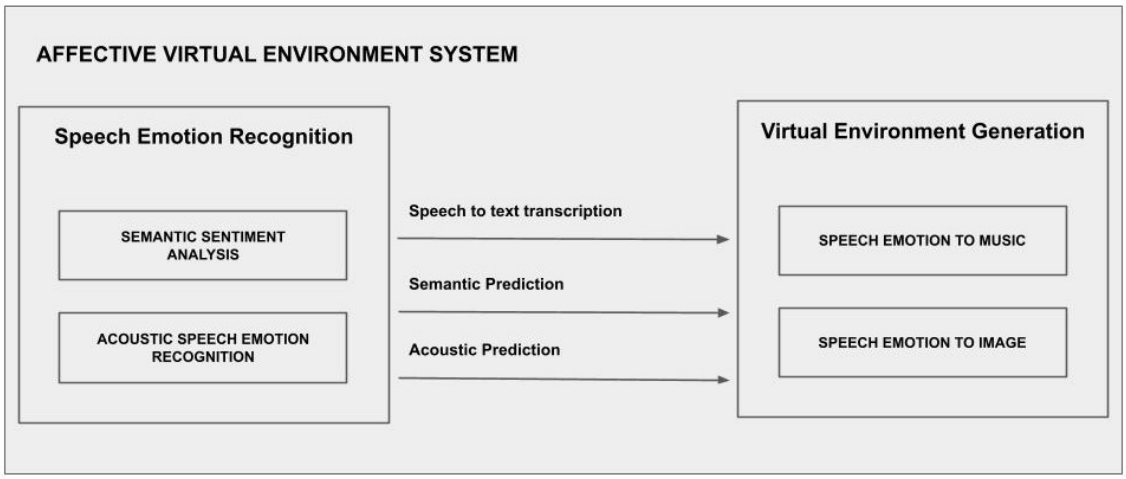


Figure 3.2: Detailed Architecture of the Affective Virtual Environment System. This diagram illustrates the dual-stream processing for emotional mapping within the AVE System.

This selection captures the progression of state-of-the-art emotion recognition techniques, ranging from classical vector-space methods to deep learning and transformer-based models. These models were employed throughout the iterative creative process and implemented across several software applications (see Chapter 7). In this study, we present the results achieved by these classifiers across both modalities. The workflow in each case follows a structured sequence of five stages:

1. **Definition of the emotional reference system:** Establishing the categorical or dimensional framework used to represent affect.
2. **Database selection:** Choosing datasets that are consistent with the defined emotional reference system.
3. **Data preprocessing:** Transforming raw data into an optimal format for analysis, including cleaning, normalization, and segmentation.
4. **Feature extraction:** Identifying and computing relevant semantic or acoustic features from the preprocessed data.
5. **Model training and evaluation:** Training classifiers and assessing their performance using metrics such as accuracy, precision, recall, and F1-score.

The recognition component outputs a textual transcription of the speech captured by the microphone together with two categorical emotional predictions, one derived from the semantic model and one from the acoustic model (see Figure 3.2).

3.1.1 Sentiment Analysis

Sentiment analysis was performed on the Round 1 split of the DynaSent benchmark (Potts et al. 2021), a corpus of English-language sentences drawn from the Yelp Academic Dataset and annotated into three sentiment categories: negative, neutral, and positive. The original training partition is substantially imbalanced, comprising 21,391 positive, 45,076 neutral, and 14,021 negative instances. To ensure unbiased estimation of model performance across classes, we applied a downsampling strategy that reduced all categories to the size of the minority class, yielding a balanced training subset of 14,021 samples per category. Validation and test sets were used as provided by the benchmark authors, each comprising 1,200 balanced samples per class (3,600 total), facilitating direct comparison with prior work (Potts et al. 2021).

Text Preprocessing

A structured normalization pipeline was applied consistently across all model families. Raw text was lowercased and stripped of punctuation, special characters, and redundant whitespace. Stopword removal was performed using the NLTK stopword list (Bird, Klein, and Loper 2009), with a targeted retention strategy: negations and contrastive conjunctions (e.g., *not*, *never*, *but*, *however*) were explicitly preserved, given their well-documented role in polarity inversion and sentiment composition (Wilson, Wiebe, and Hoffmann 2005; Taboada et al. 2011). Morphological normalization was subsequently performed via lemmatization using the `spaCy` library (Honnibal et al. 2020), mapping inflected forms to their canonical base representations through efficient batch processing via `nlp.pipe`. This full pipeline reduced average sentence length from 14 to 7.3 tokens and contracted the vocabulary from 48,712 to 20,049 unique lemma types.

To isolate the contribution of preprocessing to classification performance, three textual representations were retained throughout the Rocchio and LSTM experiments: the raw original text, a cleaned (but non-lemmatized) version, and a fully lemmatized version. The impact of each representation was treated as an empirical question and evaluated independently rather than assumed to be uniformly beneficial.

Feature Extraction and Model Families

Three model families of increasing representational complexity were evaluated. For the Rocchio classifier, text was vectorized using TF-IDF with unigram and bigram features, sublinear term-frequency scaling, and a minimum document frequency filter to suppress sparse terms. For LSTM-based classifiers, token sequences were converted to integer indices using a vocabulary of 24,093 tokens; out-of-vocabulary tokens were mapped to a reserved `<OOV>` symbol (observed OOV rate on the test set: 3.7%) and sequences were padded to a uniform length of 128 tokens. Two embedding strategies were compared: 300-dimensional pretrained GloVe vectors (`glove.840B.300d`) and randomly initialized embeddings learned from scratch. For BERT-based classifiers, input sequences were encoded using the `bert-base-uncased` tokenizer, including the insertion of `[CLS]` and `[SEP]` special tokens, with padding and truncation applied to a maximum length of

128 tokens. Attention masks were generated to distinguish content tokens from padding, and the resulting tensors were organized into `tf.data.Dataset` pipelines with a batch size of 64.

Training Protocol and Evaluation

Evaluation followed a hold-out validation strategy using the predefined DynaSent benchmark splits rather than k-fold cross-validation. This approach was selected for two reasons: first, the benchmark's balanced test partitions are designed to reflect realistic sentiment distributions and to enable reproducible comparison across studies; second, the large training corpus provides sufficient statistical mass to yield reliable parameter estimates without the computational overhead of iterative resampling. LSTM results were averaged over three independent runs with distinct random seeds to control for initialization variance. BERT models were fine-tuned for six epochs with validation monitoring; the Adam optimizer was used throughout, with learning rate 1×10^{-5} and epsilon 1×10^{-8} for transformer models, and learning rate 1×10^{-3} for recurrent models. Early stopping with patience of three epochs was applied to LSTM training.

3.1.2 Speech Emotion Recognition

Acoustic emotion classification was performed on the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (Livingstone and Russo 2018), a high-quality acted corpus comprising 1,440 speech recordings from 24 professional actors (12 male, 12 female). Each actor produces two semantically neutral sentences—*"Kids are talking by the door"* and *"Dogs are sitting by the door"*—across eight emotional states (neutral, calm, happy, sad, angry, fearful, disgust, and surprised) at two intensity levels (normal and strong). The controlled lexical content of the stimuli ensures that emotional information is conveyed exclusively through prosodic, spectral, and voice-quality characteristics rather than through lexical semantics, rendering RAVDESS particularly suitable for the acoustic-only recognition task pursued here. The full speech-only subset was used without downsampling. The class distribution is near-uniform at 192 samples per category, with the sole exception of the neutral class (96 samples); class weights were applied during training to compensate for this asymmetry. Audio files are approximately 3–4 seconds in duration, recorded at 48 kHz in WAV format.

Acoustic Feature Extraction

Acoustic features were extracted using the openSMILE toolkit (Eyben, Wöllmer, and Schuller 2010), configured with the Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPSv02) at the *Functionals* level, yielding a standardized 88-dimensional feature vector per utterance. The eGeMAPS feature set was designed specifically for paralinguistic and affective computing research, providing a compact but expressive representation of the acoustic correlates of emotion (Eyben et al. 2016). Functionals summarize frame-level low-level descriptors (LLDs) over each utterance using statistical operators—including mean, standard deviation, and percentile estimates—thereby converting time-varying signals into a fixed-length representation amenable to

utterance-level classification. The selected feature dimensions span five complementary acoustic domains: prosodic cues (fundamental frequency, loudness), spectral descriptors (MFCCs, spectral flux, spectral balance measures), voice-quality correlates (jitter, shimmer, harmonics-to-noise ratio), formant characteristics (F1–F3 frequency, bandwidth, amplitude), and temporal dynamics (voiced/unvoiced segment statistics). As a comparative condition, a reduced 40-dimensional representation based exclusively on Mel-Frequency Cepstral Coefficients (MFCCs) was also evaluated to isolate the contribution of cepstral features to classification performance.

Data Augmentation

To enhance model generalization and mitigate the effects of limited corpus size, waveform-level data augmentation was applied to the RAVDESS recordings prior to feature extraction. Four complementary strategies were implemented: (i) **white noise injection**, in which zero-mean Gaussian noise scaled to the signal's standard deviation was additively superimposed on the waveform to simulate environmental acoustic variability; (ii) **time stretching**, which modifies speaking rate without altering pitch by a factor of 0.8 to simulate inter-speaker tempo differences; (iii) **time shifting**, which applies circular temporal displacement to simulate onset variability; and (iv) **pitch shifting** by 0.7 semitones, which introduces speaker-level F0 variation without altering utterance duration. All augmented waveforms were subsequently processed through the identical eGeMAPS extraction pipeline applied to original recordings. This augmentation scheme expanded the effective training corpus from 1,440 to 4,560 utterances—a threefold increase that was designed to improve the robustness of convolutional and recurrent models to acoustic variability.

Model Architectures and Training Protocol

Three neural architectures of increasing complexity were evaluated. The Multilayer Perceptron (MLP) comprised two hidden layers (64 and 32 units) with ReLU activations and dropout regularization, followed by a softmax output layer. The 1D Convolutional Neural Network (CNN) consisted of four convolutional blocks with filter sizes of 128, 256, 128, and 64 (kernel size 5, ReLU activation), each followed by batch normalization and max-pooling, with dropout applied after the third convolutional block and in the dense layers. The CNN+LSTM hybrid extended this convolutional backbone with a unidirectional LSTM layer of 64 units interposed before the dense classification head, enabling temporal integration across the ordered sequence of acoustic descriptors extracted by the convolutional filters.

All models were compiled using the Adam optimizer and categorical cross-entropy loss, and trained for up to 150 epochs with batch size 10. Early stopping (patience = 10 epochs) monitored validation loss and restored the best-performing weights upon termination. A learning rate scheduler was applied to facilitate stable convergence. MLP experiments were conducted on both the original and MFCC-reduced datasets; CNN and CNN+LSTM experiments were conducted on both the original and augmented eGeMAPS datasets to isolate the effect of augmentation on convolutional and hybrid architectures. The model's generalization capability was assessed using

a single-split hold-out method. Performance metrics, including per-class precision, recall, and F1-score, were calculated on a final held-out test set that remained strictly isolated during the feature extraction and augmentation phases to prevent data leakage.

3.2 Assessing Irony

To assess emotional prosodic irony, we examined oppositions in emotional valence between semantic predictions derived from lexical content and acoustic predictions derived from prosodic features (see Figure 3.3). We used the IEMOCAP database (Busso et al. 2008) to gather american-english utterances labeled both acoustically and semantically. Acoustic classification was directly obtained from human annotations. Sentiment analysis was obtained using a fine-tuned BERT model on DynaSent data, with balanced training and weighted fine-tuning.

We selected the following ironic configurations:

Figure of Language	**Semantic Sentiment**	**Acoustic Prosody**
Kind Irony	Negative	Happy
Sarcasm	Positive	Angry
Sincere (anger)	Negative	Angry
Sincere (happy)	Positive	Happy

Table 3.1: Semantic-Acoustic opposition in ironic configurations (Kind Irony and Sarcasm) vs. the sincere expressions in literal speech.

The two ironic configurations, Kind Irony and Sarcasm, are defined by a valence reversal between channels: Kind Irony pairs negative semantic content with happy prosody, while Sarcasm pairs positive semantic content with angry prosody. The two sincere configurations serve as matched controls in which semantic and acoustic valence are aligned. Anger was selected as the prosodic anchor for the Sarcasm condition in particular, following Bryant & Fox Tree 2005, who demonstrated empirically that listeners systematically conflate sarcastic and angry vocal expressions, making anger the most theoretically motivated and methodologically demanding point of contrast.

Sentiment analysis was executed by fine-tuning a pre-trained BERT-Based model on the DynaSent dataset containing textual samples labeled as negative, neutral, or positive. The model was first trained on a balanced subset from Round 1 (14,021 samples per class) using an Adam optimizer with a learning rate of 2×10^{-6} and sparse categorical cross-entropy loss, trained for five epochs. It was then fine-tuned for three epochs on a balanced subset from Round 2 obtained through undersampling (approximately 2,448 samples per class), while also applying class weights during training to mitigate residual class imbalance. The model achieved 66% accuracy on the Round 1 test set, with macro and weighted F1 scores of 0.66, and class-specific F1 scores of 0.65 (negative), 0.70 (neutral), and 0.62 (positive). For Round 2, it reached 67% accuracy and

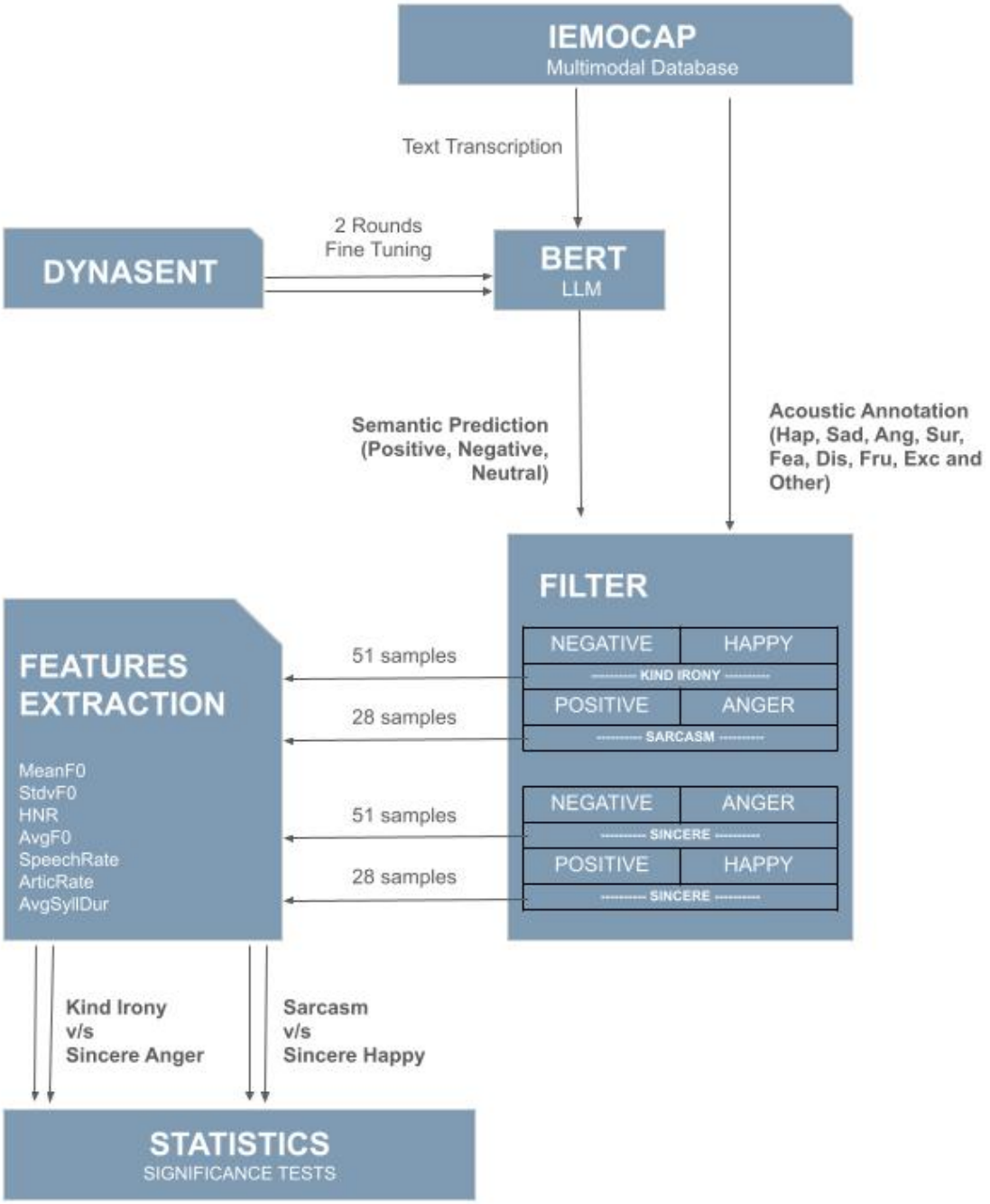


Figure 3.3: Experimental pipeline for ironic sample filtering and analysis. This architecture details the multimodal processing of emotional nuances, specifically focusing on the distinction between sincere anger vs. kind-ironic speech, and sincere happy vs. sarcastic speech.

macro and weighted F1 scores of 0.66, with class-specific scores of 0.64 (negative), 0.72 (neutral), and 0.63 (positive).

We obtained 51 samples of Kind Irony and 28 samples of Sarcasm, which were contrasted with 51 samples each of sincere angry and sincere happy speech. From these samples, we extracted seven prosodic features commonly associated with affective and expressive speech (see 3.2).

Feature	**Description**	**Units**
Mean F0	Average fundamental frequency of the speech signal	Hz
F0 Standard Deviation	Variability of the fundamental frequency	Hz
HNR	Harmonics-to-noise ratio, indicating voice quality	dB
Mean Loudness	Average perceived intensity of the speech signal	dB
Speech Rate	Number of syllables produced per second	syllables/s
Articulation Rate	Speech rate excluding pauses	syllables/s
Mean Syllable Duration	Average duration of syllables in continuous speech	ms

Table 3.2: Prosodic features extracted for statistical analysis, including descriptions and measurement units.

For each prosodic feature, we conducted statistical analyses to identify significant differences between ironic speech (Kind Irony and Sarcasm) and their sincere counterparts. Prosodic feature extraction was performed using Praat Boersma and Weenink 2001.

The statistical analysis pipeline included:

- **Shapiro–Wilk test** to assess normality of feature distributions. The test statistic W ranges from 0 to 1, where values close to 1 indicate that the distribution does not deviate significantly from normality, and values approaching 0 indicate strong departures from it. We used a significance threshold of $\alpha = 0.05$, below which the normality assumption is rejected and non-parametric alternatives are adopted.

- **Levene's test** to evaluate homogeneity of variances across groups. The resulting p-value indicates whether the variance of the two compared groups can be considered equal; values below $\alpha = 0.05$ indicate heteroscedasticity, informing the choice of test variant applied subsequently.

- **Student's *t*-test or Mann–Whitney *U* test**, selected according to distributional assumptions. The t-test is applied when normality and homogeneity of variance hold; otherwise, the non-parametric Mann–Whitney U test is used. In both cases, we report two-tailed p-values, where values below $\alpha = 0.05$ are considered statistically significant. We expect ironic speech conditions to differ significantly from their sincere counterparts on at least a subset of prosodic features, consistent with the hypothesis that irony introduces measurable acoustic disruption relative to aligned emotional baselines.

- **Principal Component Analysis (PCA)** for exploratory dimensionality reduction and feature structure analysis Jolliffe and Cadima 2016. Each principal component explains a proportion of the total variance in the feature space, ranging from 0 to 1, where higher values

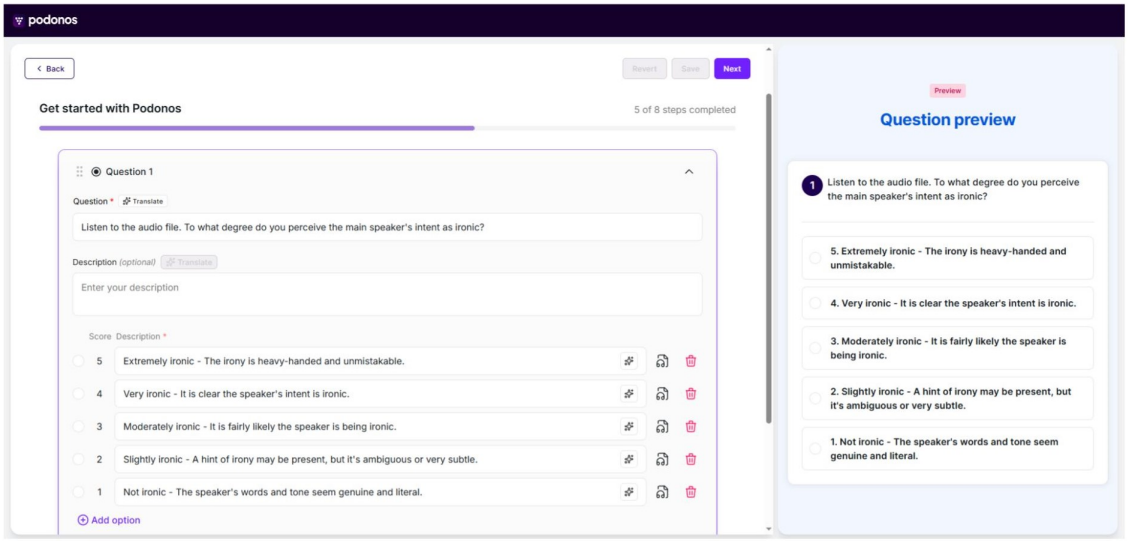


Figure 3.4: PODONOS web interface for perceptual audio analysis, used to upload and process speech samples.

indicate greater explanatory power. We expect the first components to capture the primary axes of prosodic variation that distinguish ironic from sincere speech, thereby revealing whether the two ironic configurations occupy distinct or overlapping regions of the acoustic feature space.

3.2.1 Perceptual Evaluation

We conducted a perceptual evaluation using the Podonos platform[4], in which a single-audio perceptual evaluation was designed to measure perceived irony (see Figure 3.4).

For each stimulus, listeners were asked the following question:

"To what degree do you perceive the main speaker's intent as ironic?"

Responses were provided on a 5-point Likert scale, where lower values indicated low perceived irony and higher values indicated high perceived irony. Six valid evaluators participated in the study, all native speakers of American English residing in the United States. The evaluator group consisted of four female and two male participants yielding a total of 312 valid ratings (52 stimuli $\times$ 6 evaluators). To avoid pseudo replication, ratings were first aggregated at the stimulus level by computing the mean perceived irony score for each audio file. All statistical analyses were performed on these stimulus-level means, treating each stimulus as an independent observational unit.

Because the ratings were collected on an ordinal Likert scale and normality could not be assumed, non-parametric statistical tests were employed. For each contrast between ironic and literal

4. Podonos is a web-based platform designed for perceptual evaluation of audio stimuli, supporting controlled listening tests and subjective assessments in audio and speech research. Available at https://podonos.org.

speech, a Mann–Whitney U test was used to assess differences in perceived irony between categories. Effect sizes were quantified using Cliff's delta (δ), which estimates the probability that a randomly selected stimulus from one category receives a higher rating than a randomly selected stimulus from the comparison category. Confidence intervals for δ were estimated using bootstrap resampling (10,000 iterations). Holm–Bonferroni correction was applied to control for multiple comparisons.

3.3 Affective Virtual Environments Generation

To generate affective virtual environments, we defined and validated a strategy for composing semantic prompts derived from the transcribed text and the bimodal emotional predictions (See Figure 3.5). Emotional profiles were inferred through statistical analysis of audiovisual features extracted from two annotated databases: one with musical excerpts and one digital images. Two prompting strategies were evaluated for text-guided AI generation of both music and images:

- **High-level prompting:** a single affective label specifies the target emotion (e.g., *"Generate sad background music"*).
- **Mid-level prompting:** structured descriptors derived from the feature profiling analysis are used instead (e.g., *"Generate music in a minor mode with high dissonance and increased harmonic instability"*).

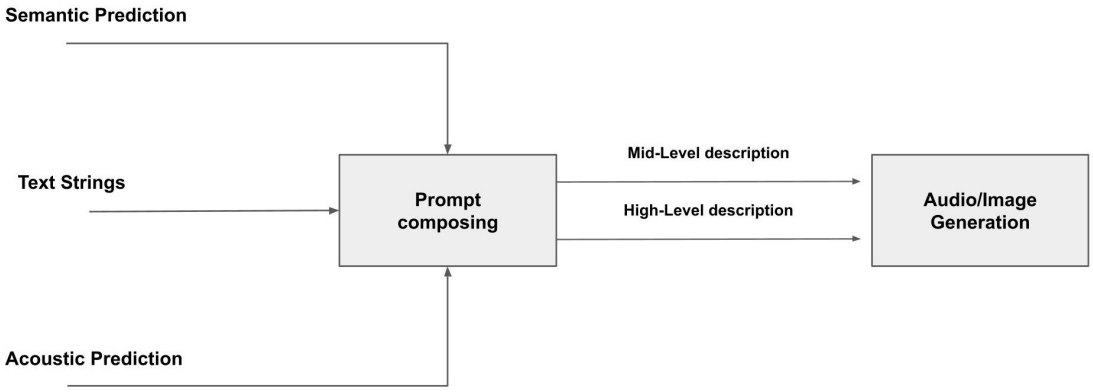


Figure 3.5: Mapping strategy for affective asset generation. The diagram depicts the transformation of multimodal emotional predictions into descriptive prompts for the synchronized synthesis of virtual environment components.

All prompts were constrained to a maximum of 77 tokens to comply with the input limitations of the generative models. We explored environments in which musical and visual emotional profiles were intentionally opposed, combining *happy* and *anger* profiles as an audiovisual analogue of verbal irony. This deliberate affective mismatch mirrors the semantic–prosodic divergence used to operationalize irony in Chapter 5.

3.3.1 Music-Based Emotional Profiling

To construct musically grounded prompts, we analyzed audio features extracted from the 4Q Audio Emotion Database (Panda, Malheiro, and Paiva 2018), comprising 45-second clips annotated according to Russell's circumplex model. From the Pop/Rock subset — the largest and most affectively coherent genre category in the corpus — clips were filtered by emotional tag: *happy* (21), *sad* (20), *anger* (58), and *gentle* (22). The collection was downsampled to 20 clips per category to ensure balanced class representation.

Features were extracted using the Essentia library and organized into five musical dimensions, as summarized in Table 3.3.

Musical Dimension	**Feature**
Melodic	Pitch salience mean / st.dev
Harmonic	Key (Krumhansl scale)
	Chord change rate
	Dissonance mean / st.dev
	HPCP entropy mean / st.dev
Rhythmic	Beats per minute (BPM)
	Onset rate
Dynamic	Average loudness
Timbral	Spectral centroid mean / st.dev

Table 3.3: Audio feature set extracted with the Essentia library, organized by musical dimension. These descriptors capture pitch prominence, harmonic structure, rhythmic activity, loudness, and spectral brightness, properties widely associated with emotional expression in music.

For each continuous feature, a standardized inferential pipeline was applied. Normality was assessed per quadrant using the Shapiro–Wilk test, and homogeneity of variances was evaluated with Levene's test. Depending on the outcome of these checks, either a one-way ANOVA or a Kruskal–Wallis non-parametric test was applied. When ANOVA reached significance ($p < 0.05$), Tukey's HSD test was used for post-hoc pairwise comparisons, and effect size was quantified via η^2. To control for false discoveries across the full feature set, Benjamini–Hochberg FDR correction was applied ($\alpha = 0.05$); only features surviving this correction were retained as emotionally discriminative.

Categorical descriptors (e.g., tonal mode, key) were analyzed using a quadrant $\times$ category frequency matrix, with Pearson's chi-square test of independence used to assess distributional differences. Significant chi-square p-values were included in the global FDR correction. Results of this analysis are reported in Section 6.1.

3.3.2 Image-Based Emotional Profiling

To derive a visual emotional profile for text-guided image generation, we extracted low-level visual features from the *Emotion6 (ROI)* dataset (Peng et al. 2015), a publicly available corpus of

1,980 naturalistic images sourced from Flickr and annotated across six basic emotion categories—*anger*, *disgust*, *fear*, *joy*, *sadness*, and *surprise*—each comprising 330 samples. Annotations were obtained via Amazon Mechanical Turk with 15 independent raters per image.

A total of 726 continuous features were extracted across five visual dimensions, as summarized in Table 3.4.

Visual Dimension	**Feature**
Color	Color Histogram (512 dims)
	Dominant Colors (9 dims)
	Color Moments (9 dims)
Texture	HOG (128 dims)
	LBP (58 dims)
	Gabor Filters (4 dims)
Image Complexity	Entropy (1 dim)
Symmetry	Horizontal Symmetry (1 dim)
	Vertical Symmetry (1 dim)
Structural	Edge Density (1 dim)
	Brightness (1 dim)
	Contrast (1 dim)

Table 3.4: Visual feature set extracted from the Emotion6 dataset, organized by visual dimension. These descriptors capture chromatic distribution, textural complexity, compositional balance, and structural properties widely associated with emotional expression in images.

Each feature was statistically evaluated following the same inferential workflow established for the musical profiling: normality and variance homogeneity were assessed using the Shapiro–Wilk and Levene's tests respectively, and hypothesis testing was conducted using parametric or non-parametric procedures accordingly. A multi-group comparison across all six emotions was performed using the Kruskal–Wallis test, with Benjamini–Hochberg FDR correction ($\alpha = 0.05$) applied to control false discovery rates across the high-dimensional feature space.

In addition to the six-emotion comparison, a focused pairwise analysis contrasted *joy* and *anger* to identify the visual features most responsible for their emotional divergence—the two poles of valence relevant to the ironic audiovisual mapping. For each feature, the most appropriate pairwise test was selected based on distributional assumptions:

- If both groups were normally distributed:
  - **Student's independent t-test** (equal variances), or
  - **Welch's t-test** (unequal variances).
- If at least one distribution violated normality:
  - **Mann–Whitney U-test** (non-parametric).

Features achieving significance ($p < 0.05$) were retained as discriminative between *joy* and *anger* and used to inform mid-level prompting strategies for affective image generation. Full feature extraction details and statistical results are reported in Section 6.2.

3.4 Integration into the Generator

The affective profiles derived from music and image analysis were integrated as a strategy to map the automatic speech emotion recognition parameters into the audiovisual space. We researched the way these profiles can be used to tune prompts that transfer an emotional print. We propose using each profile deduced from the feature analysis to build ironic virtual environments. As text prompts, we tried three semantic levels to describe the emotional state of happiness/joy and anger. Following verbal irony equivalence, we have four possible combinations between the two modes of communication for the two emotions studied (see Table 7.2).

Visual Profile	Musical Profile	Audiovisual Space
Anger	Happy	Kind Irony
Anger	Anger	Sincere Anger
Joy	Anger	Sarcasm
Joy	Happy	Sincere Happiness

Table 3.5: Cross-Modal Affective Mapping Logic: Combinatorial rules for synthesizing ironic and sincere audiovisual spaces based on visual and musical emotional profiles.

We used *DeepAI* image generation through their documented API to produce image textures[5]. We generate nine versions for each prompt for each emotion (see Figure 7.26). For Music, we generated nine 30-second Audio Clips for each emotion using *musicGen*[6]. We implemented the mid-level semantic descriptions to develop a web app in p5.js, a sketch for the visualization of the four possible scenarios (see Figure 7.14).

3.4.1 Perceptual Technology Validation

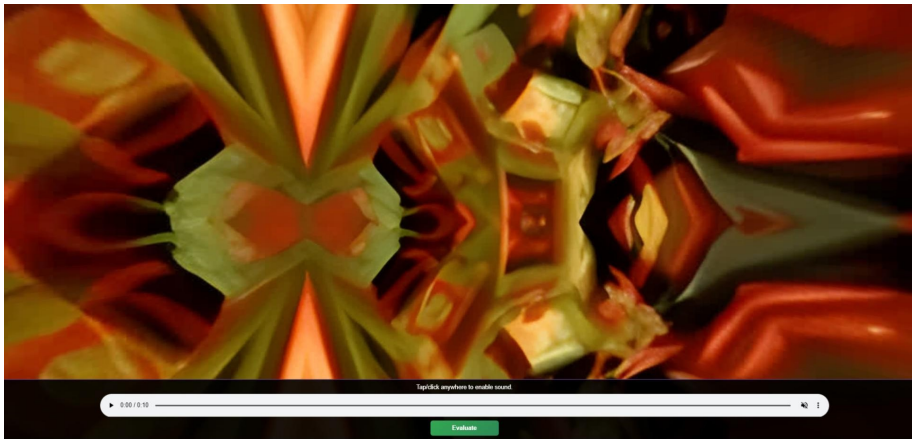
To evaluate the perceptual coherence of the generated audiovisual environments, a total of 30 virtual environments were produced. These environments comprised four affective configurations—*Kind Irony*, *Sarcasm*, *Sincere Anger*, and *Sincere Joy*—with eight environments generated for each configuration using two prompting strategies: high-level prompts and mid-level prompts. From this collection, one representative environment was selected for each combination of affective configuration and prompting strategy, resulting in a total of eight stimuli used in the perceptual evaluation.

A subjective listening and viewing experiment was conducted using an online questionnaire implemented through Google Forms (see Fig. 3.6). Participants accessed the environments through a web interface that allowed interactive exploration of the 360° panoramic scene while simultaneously listening to the corresponding generative soundtrack. After experiencing each environment, participants were asked to evaluate the degree of audiovisual coherence by answering the following question:

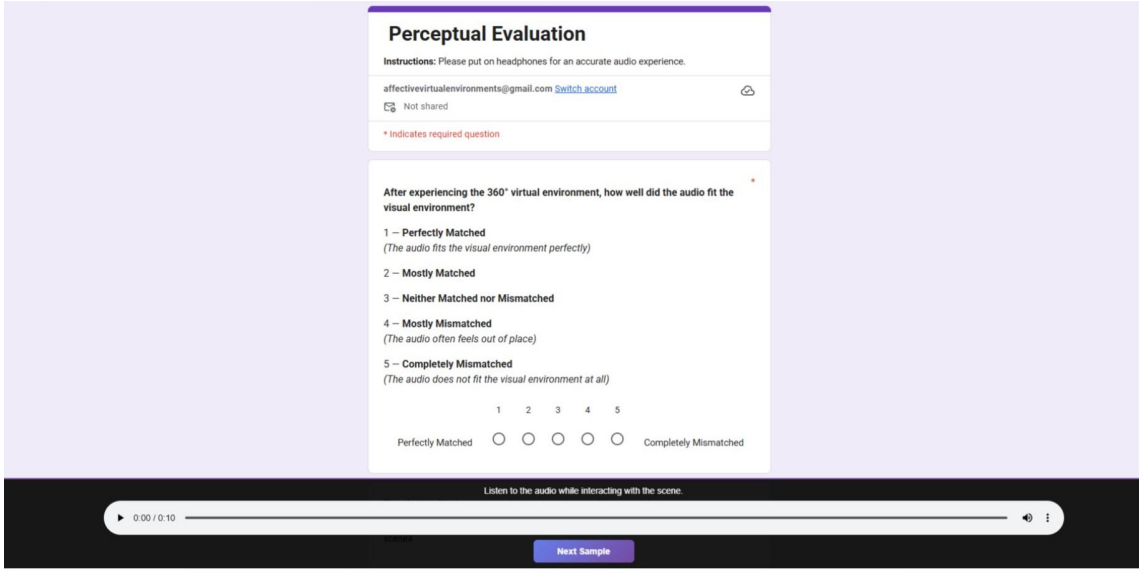
> *"After experiencing the 360° virtual environment, how well did the audio fit the visual environment?"*

5. Obtained from https://deepai.org/machine-learning-model/text2img
6. Obtained from https://colab.research.google.com/github/sanchit-gandhi/notebooks/blob/main/MusicGen.ipynb

Responses were recorded using a five-point Likert scale ranging from *Perfectly Matched* (1) to *Completely Mismatched* (5). This scale was designed to capture perceived audiovisual congruency and to assess whether environments constructed with opposing emotional modalities would produce stronger perceptions of mismatch than environments with aligned emotional profiles. A total of 30 valid ratings were collected across the eight audiovisual environments.



(a) Web platform interface for stimulus presentation.



(b) Likert scale interface for subjective congruency rating.

Figure 3.6: The web platform interface used for the perceptual evaluation of generated environments. Participants rated the alignment of audiovisual stimuli on a five-point Likert scale ranging from *Perfectly Matched* to *Completely Mismatched*.

Because the responses were collected using an ordinal Likert scale and the number of observations per condition was limited, non-parametric statistical analyses were employed. Specifically,

Mann–Whitney U tests were used to compare perceived mismatch scores between two experimental contrasts:

1. ironic environments (*Kind Irony* and *Sarcasm*) versus sincere environments (*Sincere Joy* and *Sincere Anger*); and

2. environments generated using high-level prompts versus those generated using mid-level prompts derived from statistically significant audiovisual features.

These tests evaluate whether the distributions of mismatch ratings differ significantly between experimental conditions without assuming normality of the data.

3.4.2 The Creative Experience

Throughout this research, a series of creative tools, prototypes, and exploratory systems were iteratively developed, evaluated, and refined. Some artifacts served primarily as instrumental components within the final Affective Virtual Environment (AVE) system, while others evolved into autonomous creative works presented in artistic and research contexts. Collectively, these artifacts form a structured sequence of experimental prototypes that contributed both technically and conceptually to the investigation of irony, affect, and multimodal virtual environments. The resulting trajectory can therefore be understood as an iterative experimental cycle. In each iteration, a technical or conceptual problem was explored through the design of a working system or media artifact. The outcomes of these implementations—including both successful behaviors and unforeseen limitations—provided empirical and experiential insights that informed the design of subsequent iterations.

In this sense, the creative artifacts function as *epistemic objects*: material configurations through which theoretical questions about emotion recognition, irony, and audiovisual perception become observable and analyzable. This iterative methodology also enabled the integration of heterogeneous forms of evidence. Quantitative results derived from machine learning models and statistical analysis were complemented by qualitative observations obtained through interaction, exhibition contexts, and perceptual exploration. The dialogue between these domains allowed the research to move fluidly between computational experimentation and artistic inquiry, maintaining methodological rigor while accommodating exploratory design processes typical of creative computing. By documenting this set of projects, the thesis foregrounds creative practice as a rigorous research method. Rather than resolving ambiguity, the cumulative process sustains it as a productive space where affect, technology, and meaning continually negotiate and inform one another.

The project has been extensively documented through dedicated web pages, which provide detailed insights into both the technical and practical aspects of the research. These online resources (accessible at https://emotional-machines.com/) include descriptions of system architectures, machine learning pipelines, interactive prototypes, creative outputs, and audiovisual demonstrations. The web-based documentation serves not only as a public record of the research process, but

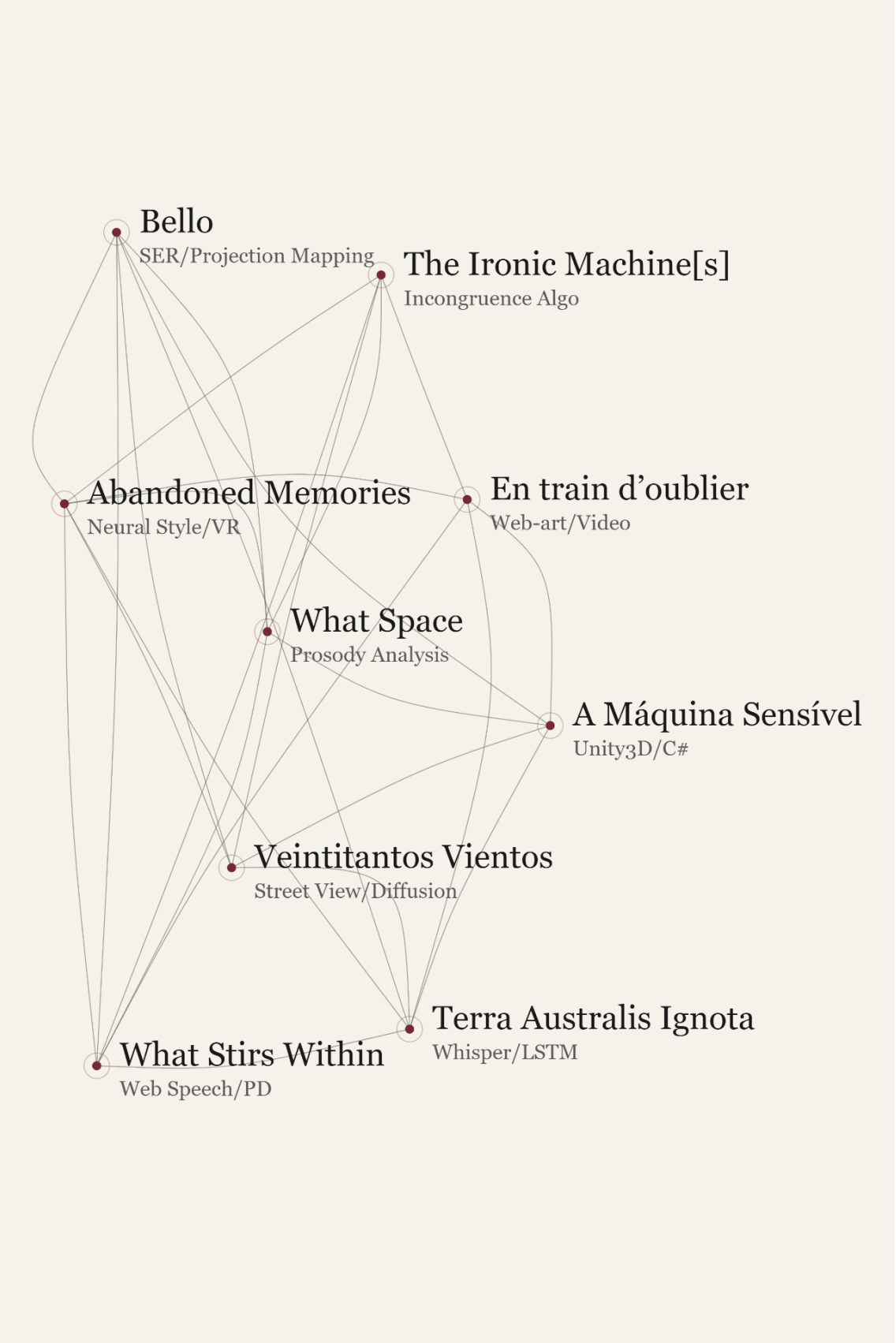


Figure 3.7: Rhizomatic network of the nine creative artifacts developed throughout the research. Each node represents an individual project, annotated with its primary technological or methodological focus. The diagram visualizes the non-linear, practice-based research process, in which successive artifacts emerge through recursive experimentation across domains such as speech analysis, machine learning, and audiovisual generation, collectively contributing to the development of the Affective Virtual Environment (AVE) framework.

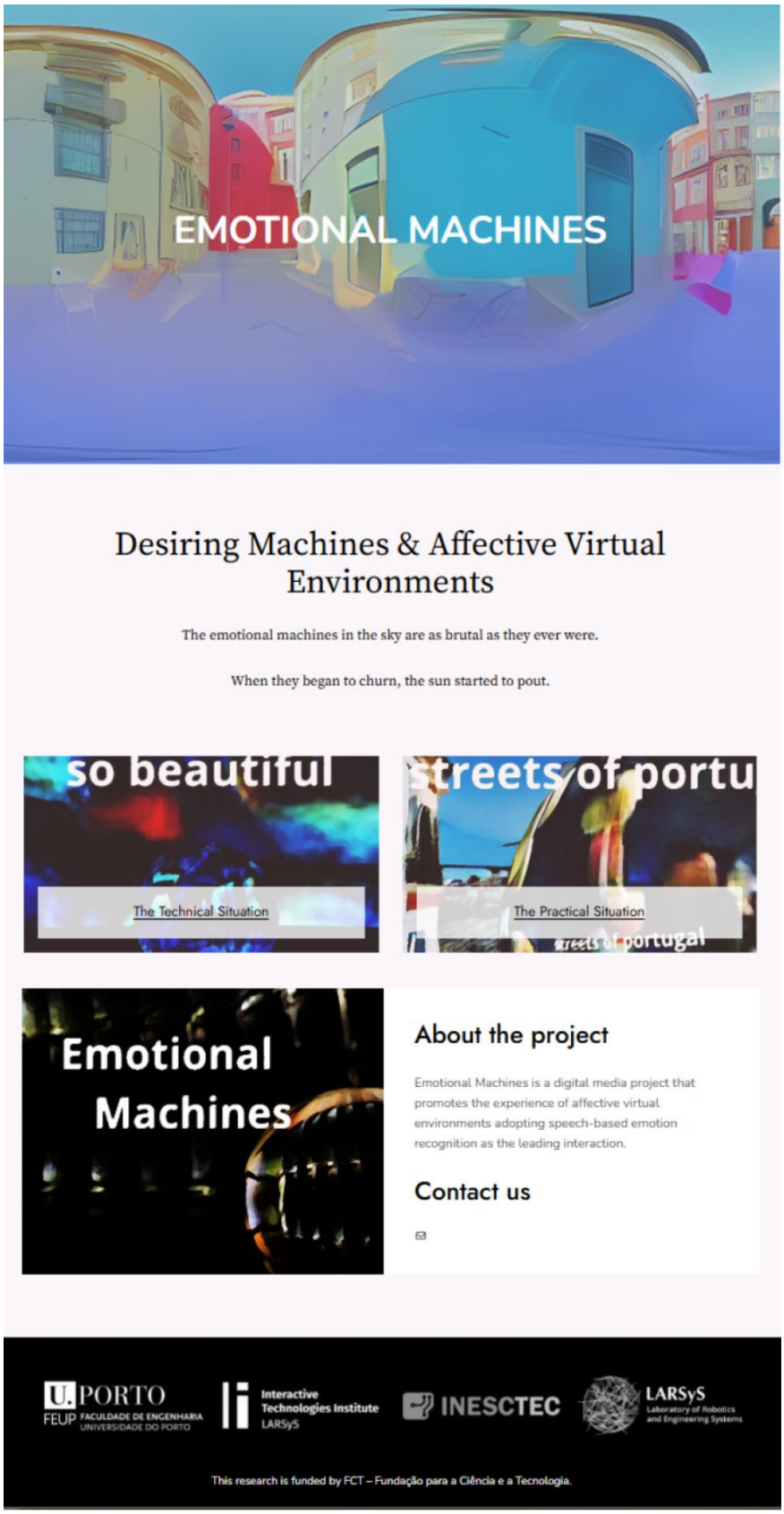


Figure 3.8: Project landing page (www.emotional-machines.com), developed and maintained during the research process. The website serves as a centralized resource for technical specifications and implementation details related to the creation of affective virtual environments.

also as a platform for sharing methodologies, design decisions, and experimental results with the broader academic and artistic community.

3.5 Ethical Considerations

In this research, participants were asked to provide voice recordings for the purposes of developing and testing the affective virtual environment system. No recordings were individualized, and all audio data were treated solely as temporarily processed files. Voice files were not stored permanently, and no identifiable information was linked to the recordings. Once processed, the files were immediately discarded to ensure the privacy and security of participants.

These measures ensure that the research complies with ethical standards for human-subject studies, maintaining confidentiality while allowing the analysis of prosodic and affective features in speech.

Chapter 4

Bimodal Speech Emotion Recognition

This chapter presents the two recognition streams developed in this thesis, designed to operate in parallel: (1) a ternary sentiment analysis pipeline that estimates the sentiment of lexical content, and (2) an acoustic speech emotion recognition pipeline that infers affective states from prosodic speech features. Together, these components constitute the core of the AVE system used in the Creative Practice (Chapter 7) and the bimodal inference mechanism introduced in Chapter 5, where verbal irony is modeled as a systematic opposition between semantic sentiment and perceived acoustic emotion. The chapter is organized into two main sections, each corresponding to one modality (text and speech), and each presenting a set of architectures that reflect increasing levels of representational complexity. The selection of these classifiers follows the broader evolution of machine learning methodologies between 2020 and 2025, ranging from feature-based approaches to deep contextual models. All models presented in this chapter have been implemented within the AVE system and deployed across the creative and analytical artifacts developed in this research.

Section 4.1 focuses on **Semantic Sentiment Analysis**. It introduces the DynaSent dataset and the text normalization pipeline used to prepare the corpus for supervised learning. Three machine learning models are evaluated. First, a prototype-based Rocchio classifier is presented as a transparent and interpretable baseline grounded in vector-space representations. Second, a recurrent neural architecture (LSTM) is examined for its ability to model sequential linguistic dependencies. Finally, a transformer-based BERT classifier is evaluated, leveraging contextualized language representations.

Section 4.2 presents the parallel pipeline for **Acoustic Speech Emotion Recognition**. After describing the RAVDESS speech corpus, the section outlines the acoustic feature extraction procedure based on the eGeMAPSv02 feature set, a standardized representation for speech emotion analysis. Three machine learning architectures are then evaluated: a feed-forward multilayer perceptron (MLP) operating on fixed-length acoustic descriptors, a convolutional neural network (CNN) designed to capture local patterns in the feature space, and a hybrid CNN+LSTM architecture that integrates convolutional feature learning with temporal modeling.

For transparency and reproducibility, the full implementation of the experimental pipelines described in this chapter is publicly available, including all preprocessing procedures, model ar-

chitectures, training configurations, and visualizations. The corresponding Colab notebooks can be accessed here:

- Sentiment Analysis Notebook
- Speech Emotion Recognition Notebook

4.1 Semantic Sentiment Analysis

As detailed in the Methodology, this study employs the DynaSent Round 1 benchmark (Potts et al. 2021) for ternary sentiment classification. The original corpus contains 48,712 unique surface tokens, with sentences averaging 14 words and a maximum length of 258 words. Following the preprocessing pipeline described in Section 3.1.1, including lowercasing, punctuation removal, selective stopword filtering, and spaCy lemmatization, the cleaned dataset contracts to 20,049 unique lemma types, with an average sentence length of 7.3 words and a maximum of 130 words. Table 4.1 illustrates representative transformations from raw to cleaned to lemmatized text, highlighting the cumulative effect of each normalization stage on lexical surface form.

Original Text	**Clean Text**	**Lemma Text**	**Label**
I used to come here about once a week and my smoothies were amazing...	used come week smoothies amazing would always ...	use come week smoothie amazing would always get ...	positive
Soft shell crab was much better.	soft shell crab much better	soft shell crab much well	positive
I'll be back for all service on my vehicle.	ill back service vehicle	ill back service vehicle	positive
Kudos to that young man for admitting he made a mistake.	kudos young man admitting made mistake	kudos young man admit make mistake	positive
I haven't had the Italian dishes but haven't seen the need...	havent italian dishes but havent seen need both...	have not italian dish but have not see need both...	positive
But was the least appetizing item all night.	but least appetizing item night	but least appetizing item night	negative
And yes, it's another CAD $25 to get all your photos...	yes another cad 25 get photos experience usb stick	yes another cad 25 get photo experience usb stick	negative

Table 4.1: Representative examples of original, cleaned, and lemmatized text from the DynaSent dataset, illustrating the cumulative transformations applied at each preprocessing stage.

4.1.1 The Rocchio Classifier

The Rocchio classifier, also referred to in the literature as a nearest centroid classifier, is a prototype-based method that assigns class labels according to the proximity of a feature vector to class-specific centroids in the TF-IDF representation space (Joachims 1997). Despite its parametric simplicity, the Rocchio classifier has been shown to perform competitively on high-dimensional, sparse text data when paired with TF-IDF representations (Joachims 1997). Its reliance on centroid proximity, however, renders it susceptible to class imbalance, a known limitation that motivates the balanced training subset constructed for this study.

4.1.1.1 Results

Table 4.2 reports test-set performance across the three variants. The cleaned representation yielded the highest overall accuracy (0.595), macro F1-score (0.593), and weighted F1-score (0.593), with the neutral class achieving its strongest individual F1 of 0.623. This outcome is consistent with prior observations that semantically ambiguous neutral sentences benefit from noise reduction without requiring full morphological normalization (Wilson, Wiebe, and Hoffmann 2005). The lemmatized variant produced comparable accuracy (0.588) but exhibited a notable decline in negative-class recall (0.438 versus 0.560 for cleaned text), suggesting that morphological normalization may inadvertently suppress surface-level cues—such as contrastive or emphatic lexical items—that are diagnostically relevant to negative polarity. Conversely, the original unprocessed text maintained the highest negative-class recall (0.552), indicating that raw lexical variation retains discriminative signal for polarity detection that cleaning partially attenuates. Taken together, these results suggest that moderate preprocessing, specifically text normalization without aggressive morphological reduction, provides the most favorable trade-off between lexical compactness and semantic retention for prototype-based classification.

Version	Acc.	P(0)	R(0)	F1(0)	P(1)	R(1)	F1(1)	P(2)	R(2)	F1(2)
Cleaned	0.595	0.582	0.560	0.571	0.553	0.714	0.623	0.686	0.513	0.587
Original	0.585	0.533	0.552	0.542	0.579	0.639	0.608	0.621	0.535	0.575
Lemma	0.588	0.586	0.438	0.501	0.522	0.746	0.614	0.645	0.532	0.583

Table 4.2: Test-set evaluation of Rocchio classifiers on DynaSent Round 1. Class 0 = negative, 1 = neutral, 2 = positive. All figures report precision (P), recall (R), and F1-score (F1) per class, together with overall accuracy (Acc.).

4.1.2 LSTM Classifier

To model sequential dependencies in sentiment-bearing text, Long Short-Term Memory networks (Hochreiter and Schmidhuber 1997) were implemented using TensorFlow and Keras (Abadi et al. 2016; Chollet et al. 2015). Input sequences were tokenized using a vocabulary of the 24,093 most frequent terms and post-padded to a uniform length of 128 tokens to ensure consistent input dimensions. The architecture comprises an embedding layer, a spatial dropout layer ($p = 0.4$)

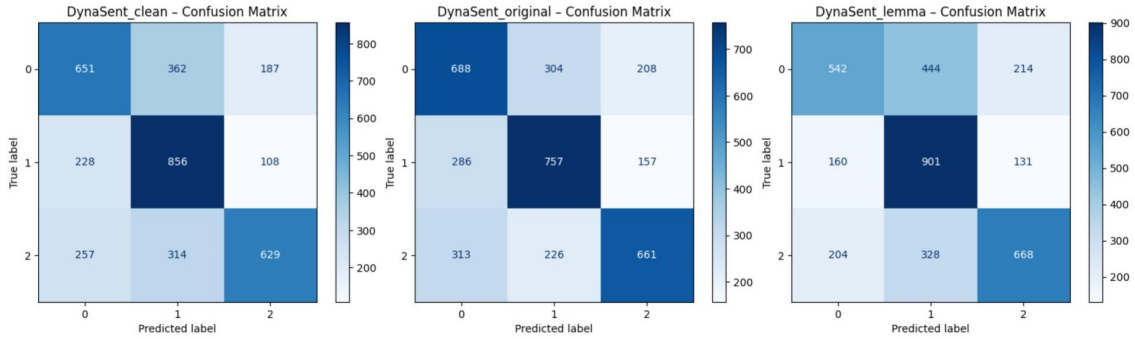


Figure 4.1: Confusion matrices for the Rocchio Classifier applied to DynaSent Round 1. Labels 0, 1, and 2 represent negative, neutral, and positive sentiment, respectively. The matrices illustrate that while the Cleaned variant (left) achieves the best overall balance, the Lemmatized version (right) exhibits superior recall for the neutral class at the expense of negative class sensitivity.

applied across entire embedding dimensions to reduce co-adaptation of feature detectors, a single LSTM layer with 196 units (dropout = 0.2, recurrent dropout = 0.2), and a softmax-activated dense output layer projecting to the three sentiment classes. Two embedding initialization strategies were compared: pretrained 300-dimensional GloVe vectors[1] with fine-tuning enabled, and randomly initialized embeddings learned exclusively from training data, in order to isolate the contribution of prior semantic knowledge to classification performance.

Init	**Acc.**	**P(0)**	**R(0)**	**F1(0)**	**P(1)**	**R(1)**	**F1(1)**	**P(2)**	**R(2)**	**F1(2)**
GloVe	0.694	0.724	0.608	0.661	0.705	0.722	0.714	0.661	0.751	0.703
Random	0.606	0.505	0.784	0.614	0.666	0.532	0.591	0.772	0.500	0.607

Table 4.3: Test-set evaluation of LSTM models on DynaSent Round 1. Class 0 = negative, 1 = neutral, 2 = positive. Results are averaged across three independent runs with distinct random seeds.

4.1.2.1 Results

Table 4.3 reports test-set performance averaged across three independent runs. GloVe-initialized embeddings substantially outperformed random initialization, achieving an accuracy of 0.694 against 0.606, a relative improvement of 14.5%. The pretrained variant produced balanced F1-scores across all three classes (0.661 negative, 0.714 neutral, 0.703 positive), indicating stable generalization without systematic bias toward any particular sentiment category. The randomly initialized model, by contrast, exhibited markedly asymmetric behavior: while it achieved the highest negative-class recall of either condition (0.784), this came at the cost of severely depressed recall for the neutral (0.532) and positive (0.500) categories, reflecting an optimization

1. Pennington, Socher, and Manning 2014. GloVe: Global Vectors for Word Representation. https://nlp.stanford.edu/projects/glove/.

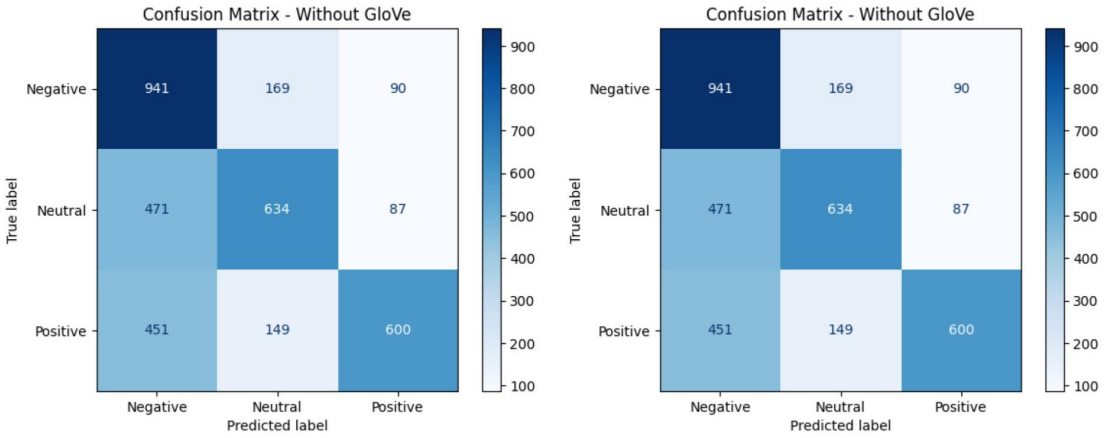


Figure 4.2: Confusion matrices for the LSTM model using randomly initialized embeddings (Without GloVe). The results illustrate the classification asymmetry noted in Table 4.3, characterized by high negative-class recall but frequent misclassification of neutral and positive instances into the negative category.

landscape biased toward the negative class rather than genuine distributional learning. This asymmetry underscores the importance of pretrained semantic representations for ensuring consistent performance across sentiment categories, particularly for the neutral class, whose lexical boundary with adjacent categories is inherently diffuse.

4.1.3 BERT-Based Classifiers

To assess the representational ceiling achievable through contextualized language modeling, two transformer-based architectures were fine-tuned on the balanced DynaSent dataset using the Bert-base-uncased encoder from HuggingFace Transformers (Wolf et al. 2020). Full tokenization and data preparation procedures are described in Section 3.1.1. Two classifier heads were compared:

- **BERT (baseline):** The contextual representation of the `[CLS]` token, produced by the pretrained encoder, is passed directly to a softmax classification layer via the standard `TFBertForSequenceClassification` architecture (`num_labels=3`).
- **BERT++ (enhanced classifier):** An intermediate fully connected layer with ReLU activation is interposed between the `[CLS]` representation and the final softmax output, providing an additional stage of task-specific feature transformation.

4.1.3.1 Results

Table 4.4 summarizes performance on the balanced DynaSent test set. The baseline BERT model achieved an accuracy of 0.718, with F1-scores of 0.709 (negative), 0.722 (neutral), and 0.723 (positive), a notably uniform profile that contrasts sharply with the class, dependent instability

Model	Acc.	P(0)	R(0)	F1(0)	P(1)	R(1)	F1(1)	P(2)	R(2)	F1(2)
BERT	0.718	0.732	0.687	0.709	0.747	0.700	0.722	0.683	0.768	0.723
BERT++	0.728	0.742	0.698	0.719	0.754	0.711	0.732	0.705	0.775	0.738

Table 4.4: Test-set evaluation of BERT-based classifiers on DynaSent Round 1. Class 0 = negative, 1 = neutral, 2 = positive.

observed in both the Rocchio and random-embedding LSTM conditions. This balanced distribution of errors reflects the ability of contextualized representations to capture the subtle pragmatic cues that distinguish adjacent sentiment categories, particularly the neutral-positive boundary. The BERT++ architecture, through the introduction of a single task-specific projection layer, yielded consistent gains across all three classes, achieving an overall accuracy of 0.728 and a macro F1-score of 0.728. The improvement, while modest in absolute terms (approximately 1 percentage point), is consistent across classes and statistically meaningful in the context of a well-optimized baseline, suggesting that an additional stage of non-linear feature recombination facilitates more discriminative use of the encoder's contextual representations.

Comparison with State-of-the-Art

To contextualize the performance of the models evaluated in this study, it is instructive to compare them against recent state-of-the-art results in ternary sentiment classification. Transformer-based architectures, particularly large pre-trained models such as BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019), and their task-specific variants, routinely achieve accuracies exceeding 0.75–0.80 on benchmark datasets such as DynaSent (Potts et al. 2021), with some ensemble-based or task-augmented approaches surpassing 0.82. LSTM-based models with contextualized embeddings (e.g., GloVe or ELMo) typically reach 0.68–0.71, reflecting strong sequential modeling but limited cross-category contextual sensitivity.

The results presented in this study align with these trends: baseline BERT achieves 0.718, while the augmented BERT++ reaches 0.728. Although these figures are slightly below the highest reported in the literature, several factors justify the observed gap. First, model complexity and parameter scale in top-performing approaches often involve extensive pre-training, ensemble techniques, or data augmentation strategies that fall outside the scope of this study, which prioritized architectural clarity and single-modal text-based evaluation. Second, certain high-performing models leverage additional contextual signals such as syntactic parse features, external lexicons, or cross-lingual transfer, which were deliberately excluded here to maintain a controlled comparison across model families.

Classical models, such as the Rocchio and LSTM baselines, demonstrate predictable performance relative to the literature: the Rocchio centroid method at 0.595 reflects the lower bound of interpretable, non-parametric approaches, while the LSTM with GloVe embeddings at 0.694 confirms the utility of distributional semantics in stabilizing neutral sentiment detection. These

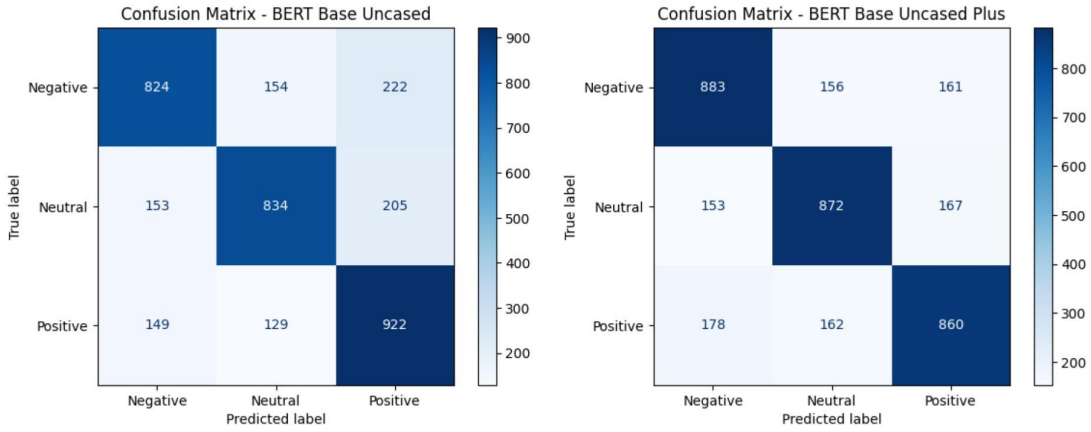


Figure 4.3: Confusion matrices for the BERT (left) and BERT++ (right) classifiers. Compared to previous models, the BERT-based architectures demonstrate significantly more balanced performance across negative, neutral, and positive classes, with the BERT++ variant showing the most robust diagonal alignment.

results are consistent with prior findings indicating that sequential compositionality enhances performance for context-dependent categories, but falls short of capturing the deep contextual interactions that transformer models inherently encode.

Overall, this comparison illustrates that the performance trajectory observed in this study; Rocchio $\rightarrow$ LSTM $\rightarrow$ BERT++, follows broader trends in sentiment analysis research. It emphasizes that while single-modal, text-based models can achieve strong performance, additional improvements observed in the state of the art often rely on multi-modal inputs, ensemble architectures, or pre-training strategies that extend beyond the scope of a controlled experimental evaluation. Importantly, situating the models relative to these benchmarks clarifies both their strengths and their limitations, providing a transparent foundation for subsequent integration into multimodal irony and emotion detection pipelines.

4.2 Acoustic Speech Emotion Recognition

This section presents the experimental results of acoustic speech emotion recognition (SER) on the RAVDESS corpus. Dataset composition, feature extraction procedures, augmentation strategies, and model architectures are described in Section 3.1.2. Results are organized to reflect a progressive evaluation of modeling capacity: from a feed-forward baseline to convolutional architectures trained on original and augmented data, and finally to a hybrid recurrent model. The role of feature representation is assessed through a direct comparison between the full eGeMAPSv02 functional set and an MFCC-only baseline. Table

Feature Category	**Selected Features**
Pitch	F0semitoneFrom27.5Hz_sma3nz_amean, stddevNorm, percentiles, slopes
Loudness	loudness_sma3_amean, stddevNorm, percentiles, slopes
Spectral Flux	spectralFlux_sma3, spectralFluxV_sma3nz, spectralFluxUV_sma3nz
MFCCs	mfcc1–4_sma3 (amean, stddevNorm); mfcc1–4V_sma3nz
Jitter	jitterLocal_sma3nz_amean, stddevNorm
Shimmer	shimmerLocaldB_sma3nz_amean, stddevNorm
HNR	HNRdBACF_sma3nz_amean, stddevNorm
Formants	F1–F3 frequency, bandwidth, amplitudeLogRelF0
Spectral Balance	alphaRatioV/UV, hammarbergIndexV/UV, spectral slopes
Temporal Features	VoicedSegmentsPerSec, Mean/Stddev Voiced and Unvoiced Seg.Length
Energy	equivalentSoundLevel_dBp, loudnessPeaksPerSec

Table 4.5: Acoustic feature set extracted using openSMILE eGeMAPSv02, organized by paralinguistic domain.

4.2.1 MLP Classifier

eGeMAPS Features

Table 4.6 reports classification performance for the MLP trained on the full 88-dimensional eGeMAPS representation. The model achieved an overall test accuracy of 71.18%, establishing a strong performance ceiling for feed-forward architectures operating on fixed-length functional descriptors. Performance was markedly uneven across emotional categories: *Angry* (F1 = 0.83), *Disgust* (F1 = 0.78), and *Surprised* (F1 = 0.79) were classified with greatest accuracy, while *Neutral*(F1 = 0.56) and *Sad* (F1 = 0.57) proved substantially more difficult. This disparity is consistent with the acoustic salience hypothesis: highly expressive affective states such as anger and surprise are characterized by pronounced deviations in fundamental frequency, loudness, and spectral energy that constitute robust discriminative cues, whereas neutral and sad speech occupy a more compressed region of the acoustic feature space with subtler and more speaker-dependent variation. The results confirm that the eGeMAPS functional set provides a sufficiently informative representation for feed-forward emotion classification within a controlled, lexically neutral corpus.

Table 4.6: Classification report for MLP on RAVDESS (eGeMAPS features).

Emotion	Precision	Recall	F1-Score
Angry	0.88	0.79	0.83
Calm	0.68	0.79	0.73
Disgust	0.73	0.84	0.78
Fearful	0.65	0.62	0.63
Happy	0.68	0.72	0.70
Neutral	0.59	0.53	0.56
Sad	0.62	0.53	0.57
Surprised	0.79	0.79	0.79
Overall Accuracy		**71.18%**	

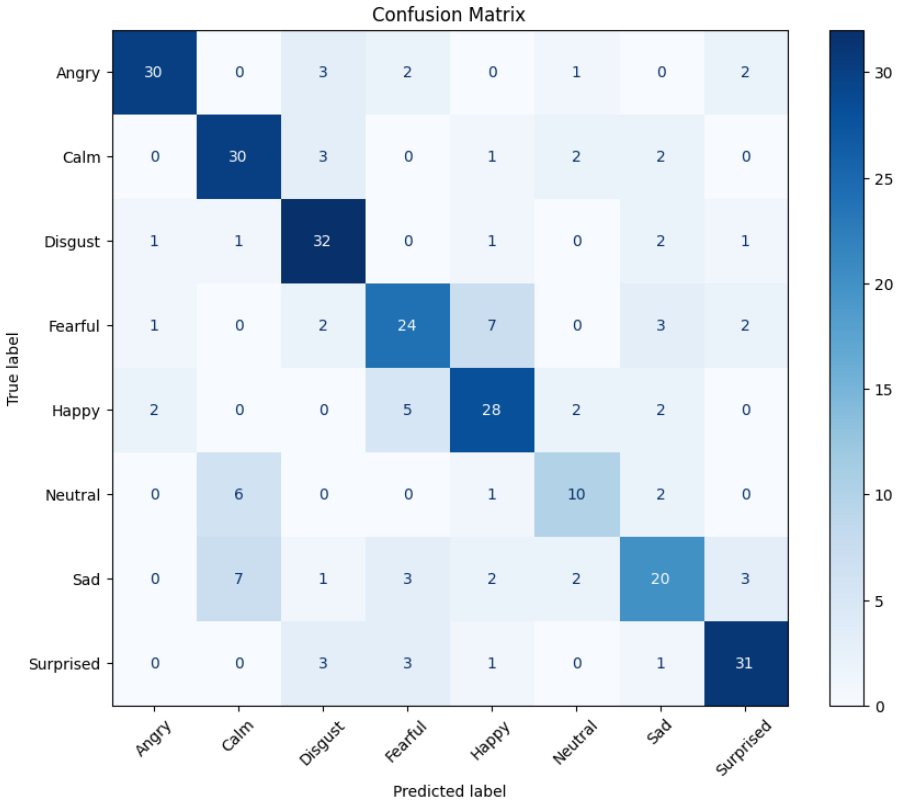


Figure 4.4: Confusion matrix for the MLP classifier trained on eGeMAPS acoustic features.

MFCC Features

To assess the contribution of the broader eGeMAPS descriptor set relative to cepstral representations alone, a second MLP was trained on 40-dimensional MFCC features under an identical training protocol. The model achieved a test accuracy of 59.38%, an absolute reduction of 11.8 percentage points relative to the eGeMAPS baseline, with markedly degraded performance across the majority of emotional categories (Table 4.7). The most pronounced declines were observed for *Happy* (F1 = 0.48 vs. 0.70) and *Neutral* (F1 = 0.47 vs. 0.56), while *Angry* remained comparatively

robust (F1 = 0.78 vs. 0.83). This pattern suggests that high-arousal emotions retain discriminability under cepstral-only representations, whereas lower-arousal and valence-ambiguous states, whose distinguishing characteristics are more heavily encoded in prosodic and voice, quality dimensions—suffer disproportionate degradation when those dimensions are absent. The result provides direct empirical support for the eGeMAPS design rationale: cross-domain acoustic coverage is not merely redundant but substantively necessary for robust eight-way emotion classification.

Table 4.7: Classification report for MLP on RAVDESS (MFCC features).

Emotion	**Precision**	**Recall**	**F1-Score**
Angry	0.81	0.76	0.78
Calm	0.58	0.76	0.66
Disgust	0.70	0.55	0.62
Fearful	0.64	0.59	0.61
Happy	0.59	0.41	0.48
Neutral	0.39	0.58	0.47
Sad	0.55	0.42	0.48
Surprised	0.50	0.67	0.57
Overall Accuracy		**59.38%**	

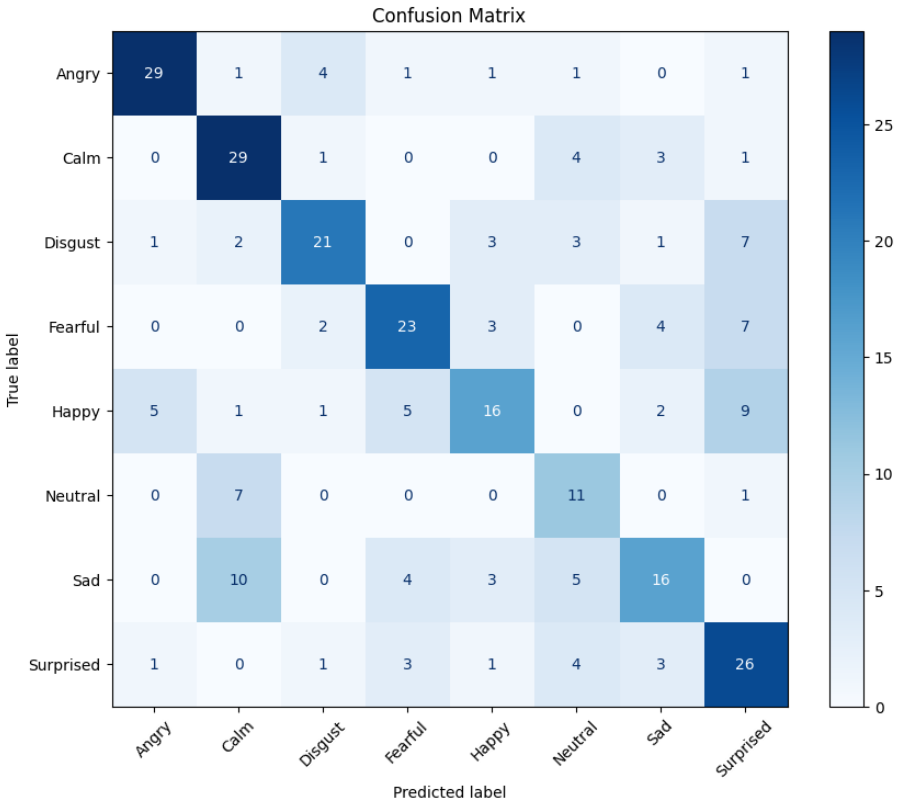


Figure 4.5: Confusion matrix for the MLP classifier trained on MFCC features.

4.2.2 CNN Classifier

Original Dataset

To evaluate the representational capacity of convolutional architectures for utterance-level SER, a 1D CNN was trained on the original eGeMAPS feature set (see Section 3.1.2 for architectural details). Despite the greater parameter count and hierarchical feature extraction afforded by convolutional operations, the model achieved a test accuracy of only 40.62%, substantially below the MLP baseline (Table 4.8). Performance was most unstable for *Sad* (F1 = 0.18) and *Happy* (F1 = 0.32), with relatively stronger but still modest results for *Calm* (F1 = 0.53) and *Surprised* (F1 = 0.53). This performance inversion—wherein a simpler architecture substantially outperforms its convolutional counterpart—can be attributed to a structural mismatch between the convolutional inductive bias and the nature of the input representation. Convolutional filters are designed to detect locally recurrent spatial or temporal patterns; applied to a fixed-length vector of utterance-level statistical functionals, they cannot exploit temporal structure because none is preserved in the representation. Furthermore, with only 1,440 training utterances, the parameter space of a four-block convolutional network is insufficiently constrained, making the model prone to overfitting—as evidenced by early stopping at epoch 22 with optimal weights obtained at epoch 12.

Table 4.8: Classification report for CNN on RAVDESS (eGeMAPS features, original dataset).

Emotion	**Precision**	**Recall**	**F1-Score**
Angry	0.52	0.34	0.41
Calm	0.50	0.55	0.53
Disgust	0.38	0.47	0.42
Fearful	0.35	0.38	0.37
Happy	0.34	0.31	0.32
Neutral	0.31	0.58	0.41
Sad	0.29	0.13	0.18
Surprised	0.50	0.56	0.53
Overall Accuracy		**40.62%**	

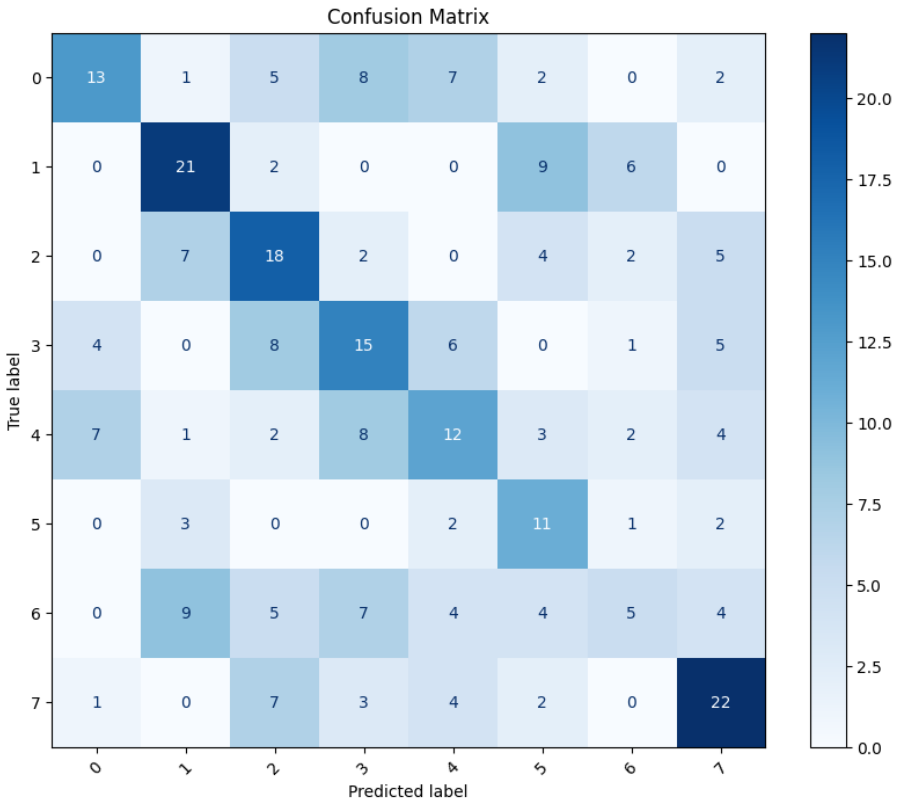


Figure 4.6: Confusion matrix for the 1D CNN classifier trained on eGeMAPS features (original dataset).

Augmented Dataset

Retraining the CNN on the augmented corpus (4,560 utterances; see Section 3.1.2) produced a substantial recovery in classification performance, raising overall accuracy from 40.62% to 63.38%, an absolute gain of 22.76 percentage points (Table 4.9). The improvement was broadly distributed across emotional categories, with the most pronounced gains for *Sad* (F1: 0.18 $\rightarrow$ 0.64), *Neutral* (F1: 0.41 $\rightarrow$ 0.66), and *Fearful* (F1: 0.37 $\rightarrow$ 0.56). Notably, the recall of *Neutral* increased from 0.58 to 0.67, while *Sad* recall improved dramatically from 0.13 to 0.74, suggesting that the acoustically subtle representations of low-arousal emotions, which were insufficiently sampled in the original corpus, benefited most from the synthetic coverage introduced by waveform-level augmentation. Strong performance was maintained or improved for high-arousal categories such as *Surprised* (F1 = 0.74) and *Happy* (F1 = 0.70). These results confirm that data scarcity, rather than architectural unsuitability per se, was the primary bottleneck for convolutional SER on RAVDESS: given sufficient training examples, the CNN recovers considerable discriminative capacity across the full emotional repertoire.

Table 4.9: Classification report for CNN on RAVDESS (eGeMAPS features, augmented dataset).

Emotion	Precision	Recall	F1-Score
Angry	0.63	0.36	0.46
Calm	0.73	0.57	0.64
Disgust	0.68	0.51	0.58
Fearful	0.49	0.64	0.56
Happy	0.68	0.72	0.70
Neutral	0.64	0.67	0.66
Sad	0.56	0.74	0.64
Surprised	0.76	0.71	0.74
Overall Accuracy		**63.38%**	

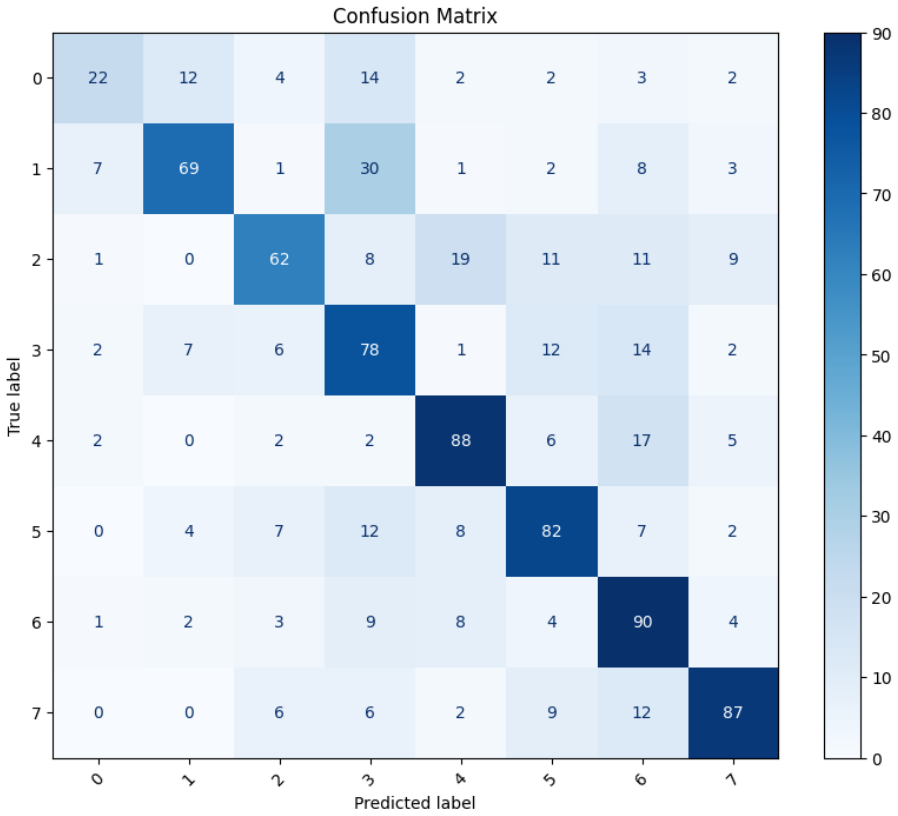


Figure 4.7: Confusion matrix for the 1D CNN classifier trained on augmented eGeMAPS features.

4.2.3 CNN+LSTM Classifier

The addition of a unidirectional LSTM layer to the convolutional backbone (see Section 3.1.2) was motivated by the hypothesis that relational dependencies among the ordered sequence of acoustic descriptors, which convolutional pooling progressively discards, may carry discriminative information for emotion classification. Training on the augmented corpus produced a test accuracy of 68.42%, representing an improvement of 5.04 percentage points over the augmented CNN and 26.8 percentage points over the non-augmented CNN baseline (Table 4.10). Per-class performance

was more uniformly distributed than in either CNN condition, with F1-scores ranging from 0.60 (*Angry*, *Fearful*) to 0.77 (*Happy*). Particularly notable is the strong performance on *Happy* (precision = 0.69, recall = 0.86), a category that had proven systematically difficult for both MLP and CNN models in the absence of sequential integration. *Calm* (F1 = 0.72) and *Surprised* (F1 = 0.72) were also classified with substantially greater precision under the hybrid architecture. The recurrent layer's contribution appears to be most pronounced for categories whose affective expression is distributed across the temporal organization of acoustic descriptors rather than localized within individual feature dimensions, a property consistent with the known prosodic complexity of happiness and calmness relative to high-arousal negative emotions. Importantly, the neutral class, which had been the most persistently problematic category across prior conditions, achieved an F1 of 0.71 under the CNN+LSTM, a result that is directly consequential for the downstream bimodal irony detection pipeline, where acoustic neutral predictions serve as a reference condition against which affective divergence is measured.

Table 4.10: Classification report for CNN+LSTM on RAVDESS (eGeMAPS features, augmented dataset).

Emotion	**Precision**	**Recall**	**F1-Score**
Angry	0.61	0.59	0.60
Calm	0.83	0.64	0.72
Disgust	0.66	0.64	0.65
Fearful	0.55	0.65	0.60
Happy	0.69	0.86	0.77
Neutral	0.75	0.68	0.71
Sad	0.64	0.72	0.67
Surprised	0.80	0.66	0.72
Overall Accuracy		**68.42%**	

Comparison with State-of-the-Art

To contextualize the present acoustic SER results on RAVDESS, it is instructive to compare the performance of the evaluated models against established benchmarks reported in the literature. Early SER systems using classical machine learning techniques—such as Support Vector Machines, Gaussian Mixture Models, or Random Forests—typically achieved 60–70% accuracy on small, controlled datasets (Schuller, Steidl, and Batliner 2009; Burkhardt et al. 2005). These approaches relied heavily on hand-engineered features and often exhibited strong speaker dependence and sensitivity to dataset characteristics.

Deep learning architectures subsequently demonstrated substantial gains. Convolutional and recurrent neural networks trained on spectrograms or low-level acoustic descriptors have reported test accuracies exceeding 80–85% on RAVDESS and similar acted corpora (Issa, Demirci, and Yazici 2020; Trigeorgis et al. 2016). Transformer-based models and pre-trained speech encoders (e.g., wav2vec 2.0, HuBERT) have further improved performance, reaching accuracies and Unweighted Average Recall (UAR) above 88% on benchmark datasets, particularly when large

amounts of labeled or augmented data are available (Pepino, Riera, and Ferrer 2021; Baevski et al. 2020).

Against this backdrop, the MLP baseline on the eGeMAPSv02 functional set (71.18%) performs within the lower range of deep learning models but achieves competitive results given the controlled, fixed-length feature representation and modest dataset size. CNN performance (63.38% with augmentation) illustrates the sensitivity of convolutional architectures to feature design and training volume: while deeper networks or spectrogram-based inputs could yield higher absolute accuracy, the use of statistical functionals constrains the exploitability of local temporal patterns. The CNN+LSTM hybrid (68.42%) improves per-class balance and robustness for low-arousal and neutral categories, reflecting an architectural adaptation that captures sequential dependencies even within fixed-length descriptors.

The decision to adopt functional descriptors rather than raw waveform or spectrogram inputs was motivated by several practical and conceptual considerations:

In summary, while the absolute accuracies of the present models do not match the highest reported in the literature, the chosen architectures and feature representations are well-aligned with the goals of controlled acoustic evaluation, computational feasibility, and downstream applicability in prosody-aware affective systems. The results therefore occupy a meaningful middle ground: competitive within classical and mid-tier deep learning SER, while retaining interpretability and practical utility for multimodal extensions.

Chapter 5

Analyzing Verbal Irony through Prosodic–Semantic Divergence

The perceived emotional content of an utterance is not uniformly distributed across communicative channels (Mehrabian and Wiener 1967). In spoken interaction, semantic meaning and prosodic delivery can operate as partially independent streams, each encoding affective information through distinct mechanisms. When these streams align, speech can be interpreted as sincere; when they decouple, the resulting semiotic incongruence becomes a valuable marker for irony (Chauhan et al. 2020; Li, Poria, and Hazarika 2024).

This chapter presents a systematic investigation of prosodic–semantic divergence as a computationally measurable phenomenon. Building on the bimodal emotion recognition framework established in Chapter 4, and consistent with Bhosale et al. (2023) and Saha et al. (2025), this work treats prosodic irony not as a discrete categorical label but as a measurable configurational state, detectable through the cross-modal opposition between lexical sentiment polarity and acoustic emotional valence. Semantic valence is derived via polar sentiment analysis applied to utterance transcriptions, while acoustic valence is operationalized through speech emotion recognition, with the *happy* and *anger* categories serving as proxies for positive and negative affective states respectively. The chapter proceeds from corpus observation through statistical inference to perceptual validation, establishing an empirical basis for proving multi-modal speech emotion recognition databases as valid source for irony analysis and recognition.

Section 5.1 introduces the IEMOCAP corpus and motivates its relevance for irony research, foregrounding the phenomenon of intra-textual emotional variation, the same lexical string annotated with distinct emotional labels across different acoustic realizations. It specifies the seven prosodic features extracted via Praat, spanning pitch level and variability, vocal intensity, temporal delivery, and voice quality. Section 5.2 presents the inferential statistical analyses and principal component analyses (PCA) used to characterize ironic speech relative to its sincere counterpart, organized by irony type and stratified by gender. Finally, Section 5.3 reports a perceptual evaluation with native speakers of American English, which provides external validation of the proposed

bimodal operationalization and confirms that the identified acoustic configurations are perceptually recoverable as ironic intent.

5.1 Data and Feature Selection

The present analysis is grounded in the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database, a widely adopted benchmark for multimodal emotion recognition research. The corpus comprises approximately 12 hours of audiovisual data collected from ten professional actors, five male and five female, engaged in dyadic scripted and improvised conversations in American english, organized into five independent sessions of two speakers each (Busso et al. 2008). Emotional annotations were performed using the ANVIL tool (Kipp 2001), employing both categorical labels and continuous attribute ratings. Crucially, annotators had access to the full communicative context, including the acoustic and visual channels as well as preceding dialogue turns, enabling a more contextually grounded emotional assessment than transcription-only paradigms permit. From the complete corpus, 10,039 samples classified across ten emotional categories were obtained. Of these, 8,068 correspond to unique textual strings, while 1,971 are repeated texts realized under differing acoustic conditions (see Figure 5.1).

A first observation from the data is that the same textual content can be perceived and annotated with different emotions depending on contextual, paralinguistic, and situational factors. For instance, the phrase "A vacation?" is consistently labeled as frustration across multiple sessions, suggesting a degree of consistency in how that phrase is delivered and perceived in certain contexts. However, other phrases like "Are you sure?" exhibit considerable variation, appearing with emotions such as sad, neutral, anger, frustration, and even happiness. This indicates that short, ambiguous utterances are highly susceptible to prosodic modulation. The case of short words or phrases like "what," "yeah," "okay," or "I see" is an illustration of language ambiguity and the complexity of human speech. The meaning of a word like "what" is overwhelmingly determined by how it is said, not by the word itself. The same sequence of phonemes can convey different emotions based on its prosody (see Figure 5.2).

These observations motivate the filtering strategy adopted in the present study. Building on the bimodal classification pipeline detailed in Chapter 3, in which a fine-tuned BERT model provides lexical sentiment predictions and IEMOCAP annotations supply prosodic emotional labels, utterances were selected on the basis of their cross-modal configuration. Specifically, we targeted the happy and anger annotation categories (representing positive/negative valence at high arousal, respectively) and crossed these acoustic labels with semantic sentiment predictions, yielding four theoretically motivated categories presented in table 5.1 .

Utterances that were semantically negative but annotated as happy were categorized as *kind irony*, while those predicted as negative and annotated as angry were labeled *sincere*. We analyzed 51 kind irony utterances (37 female, 14 male) and compared them to 51 sincere utterances (31 female, 20 male).

Text	Emotion
I don't want you to go.	sad
I know, I know. I don't want to go either bab...	sad
What's going on?	exc
Uh, -Um	sad
Why the sad face?	hap
they need me out there sooner than I thought	sad
Where? Who? What are you talking about?	sur
You went out of your way to torture me over P...	ang
No I didn't. You made up the whole thing in ...	neu
You must admit he was in love with you.	ang
Just a little perhaps but nothing serious.	neu

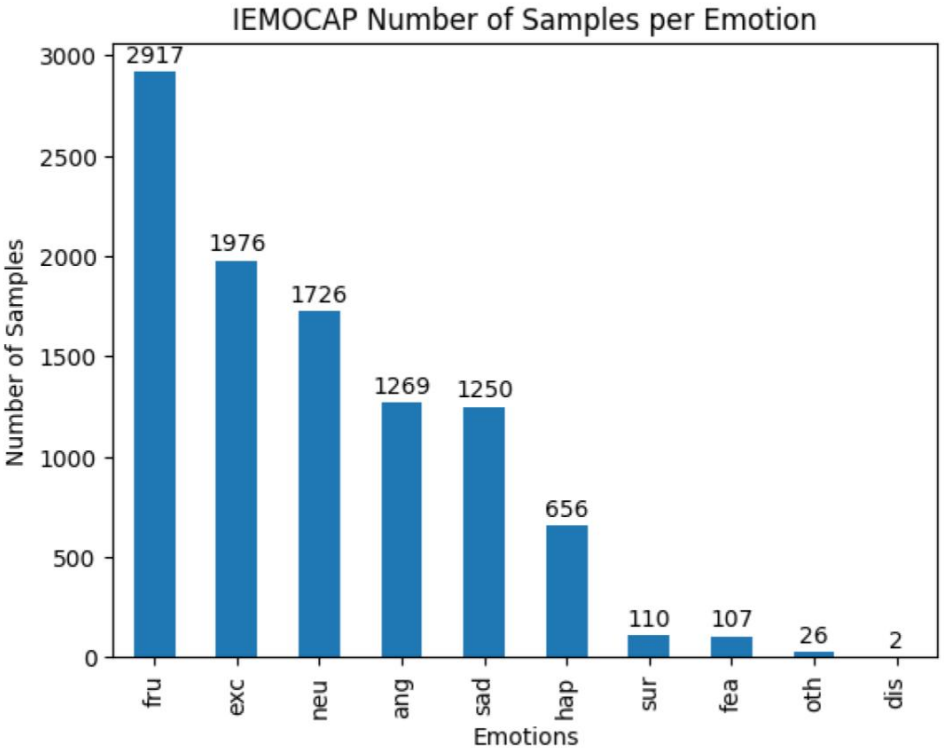


Figure 5.1: Overview of the IEMOCAP Multimodal Database. (Top) Representative text transcriptions paired with their corresponding emotional labels. (Bottom) Distribution analysis showing the total number of samples categorized per emotion across the dataset.

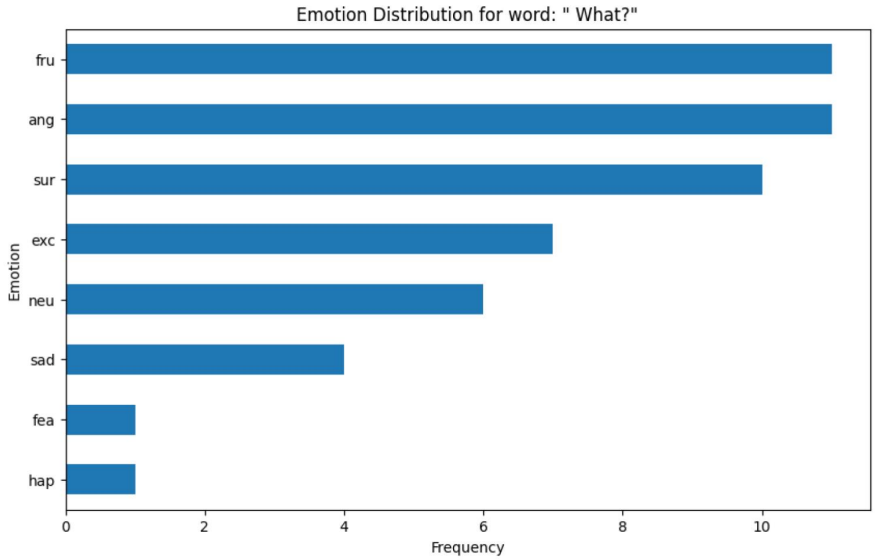


Figure 5.2: Emotion distribution for the lexical token "What?": A frequency analysis illustrating the prosodic variability of a single word across the IEMOCAP dataset.

Examples of *Kind Irony* (negative sentiment with a happy tone, i.e., praise-by-blame):

- "I'll never forgive you for that."
- "Oh, thank you, dear. The same applies to you."
- "Shh. . ."

We also obtained 28 utterances labeled as *sarcasm* (positive semantic content with an angry prosody, i.e., blame-by-praise), spoken by 6 female and 22 male actors. These were compared against 28 sincere utterances with matching positive sentiment and happy prosody (14 female, 14 male).

Examples of *Sarcasm* (positive sentiment with an angry tone):

- "Very amusing."
- "Okay, fine. I'll go over there. Thanks a lot. You've been really helpful."
- "Thanks, that helps."

To characterize the prosodic signature of each irony type, seven acoustic features were extracted from pre-filtered speech samples using Praat (Boersma and Weenink 2001). Feature selection was guided by established frameworks in affective speech research, targeting dimensions most consistently implicated in the acoustic encoding of emotional state: pitch level and variability, voice quality, vocal intensity, and temporal delivery. Consistent extraction parameters were applied across all utterances to ensure measurement comparability. The selected features and their corresponding descriptions are summarized in Table 5.2.

Category	Semantic Sentiment	Prosodic Emotion
Kind Irony	Negative	Happy
Sarcasm	Positive	Angry
Sincere (Anger)	Negative	Angry
Sincere (Happy)	Positive	Happy

Table 5.1: Bimodal configuration of semantic sentiment and prosodic emotion used to identify ironic and sincere speech categories.

Feature Name	Abbreviation	Unit	Description
Fundamental frequency mean	MeanF0	Hz	Average pitch over the sample
Fundamental frequency std. deviation	StdvF0	Hz	Pitch variability
Harmonics-to-Noise Ratio	HNR	dB	Voice clarity
Intensity Average	IntensityAvg	dB	Loudness
Speech Rate	SpeechRate	syll/s	Syllables/sec (incl. pauses)
Articulation Rate	ArticRate	syll/s	Syllables/sec (no pauses)
Average Syllable Duration	AvgSyllDur	s	Speaking time / syllables

Table 5.2: Descriptive summary of selected prosodic features and their corresponding units of measurement.

5.2 Statistical Analysis

Following the statistical analysis protocol detailed in Section 3, we evaluated distributional assumptions using Shapiro–Wilk and Levene's tests prior to selecting appropriate inferential tests (Student's t-test, Welch's t-test, or Mann–Whitney U). Principal component analysis was conducted on the seven-dimensional prosodic feature space to characterize the underlying covariance structure. Statistical analyses were implemented in Python 3.10 using SciPy, with PCA performed via Scikit-learn, ensuring full computational reproducibility.

Kind Irony vs Sincere (Anger)

The statistical and PCA results converge on a coherent picture of kind irony as a prosodic strategy organized around attenuation rather than intensification of affective expression. This pattern distinguishes it fundamentally from sincere anger, which is characterized by relatively elevated pitch and intensity, and provides the inferential baseline against which ironic deviation is measured.

Pitch and Voice Quality as Primary Carriers of Irony. Mean fundamental frequency emerged as the most robust acoustic correlate of kind irony across both genders, with statistically significant pitch lowering observed for male speakers (-4.57 semitones, $p < .01$) and female speakers

Gender	Feature	Sincere		Kind Irony		Test	P-value	Sig.
		Mean	Stdv	Mean	Stdv			
Female	meanF0Hz	274.93	45.84	239.64	39.41	Student t-test	0.001100	True
	stdevF0Hz	67.70	16.50	65.07	19.37	Student t-test	0.553417	False
	HNR	9.43	1.93	10.66	1.79	Student t-test	0.008130	True
	IntAvg	69.29	9.53	61.11	8.18	Student t-test	0.000308	True
	SpeechRte	3.34	0.95	3.66	0.67	Mann-Whitney U	0.209121	False
	ArtRte	4.41	1.01	4.45	0.68	Student t-test	0.872636	False
	AvgSylDur	0.24	0.07	0.23	0.04	Mann-Whitney U	0.990175	False
Male	meanF0Hz	190.18	48.74	146.02	58.50	Mann-Whitney U	0.006016	True
	stdevF0Hz	54.32	26.01	55.17	45.48	Mann-Whitney U	0.335899	False
	HNR	8.92	2.23	8.62	2.66	Student t-test	0.720853	False
	IntAvg	68.40	9.12	57.68	6.82	Student t-test	0.000756	True
	SpeechRte	3.38	1.19	3.02	0.83	Student t-test	0.344125	False
	ArtRte	4.28	1.18	3.98	0.83	Student t-test	0.416622	False
	AvgSylDur	0.27	0.14	0.26	0.06	Mann-Whitney U	0.201520	False

Table 5.3: Comparative Statistical Analysis of Prosodic Features. This table presents the mean and standard deviation for vocal descriptors across gender, identifying statistically significant differences ($p < 0.05$) between sincere and ironic expressions to validate the experimental filtering pipeline.

(-2.38 semitones, $p < .001$). Crucially, this pitch reduction is not accompanied by a corresponding increase in pitch variability: the standard deviation of F0 (StdvF0) does not differ significantly between conditions for either gender.

The PCA loadings reinforce this interpretation. Pitch-related features (MeanF0 and StdvF0) load substantially on the second principal component (PC2), accounting for approximately 26–27% of total variance, while PC1 is dominated by intensity and temporal features and explains over 42% of variance in both gender groups. This architecture suggests a two-tier prosodic structure: an energetic-temporal dimension (PC1) that indexes overall vocal effort and delivery rhythm, and a spectral dimension (PC2) that captures pitch-level configuration as the primary carrier of ironic register.

Intensity Reduction and Affective Containment. Intensity attenuation is observed for both male and female speakers with high statistical reliability ($p < .001$ in both cases), establishing loudness reduction as one of the most stable cross-gender correlates of kind irony. For female speakers, kind irony is additionally associated with a significant increase in harmonics-to-noise ratio (HNR; $+13.05\%$, $p < .01$), indicating a cleaner and more periodic phonatory signal relative to sincere anger. In the PCA space, HNR contributes primarily to higher-order components (PC3–PC5), implying that while voice quality does not dominate the variance structure, it constitutes a secondary discriminative dimension, particularly relevant for female ironic expression.

Temporal Stability as a Structural Signal. Temporal features (speech rate, articulation rate, and average syllable duration) do not differ significantly between conditions for either gender,

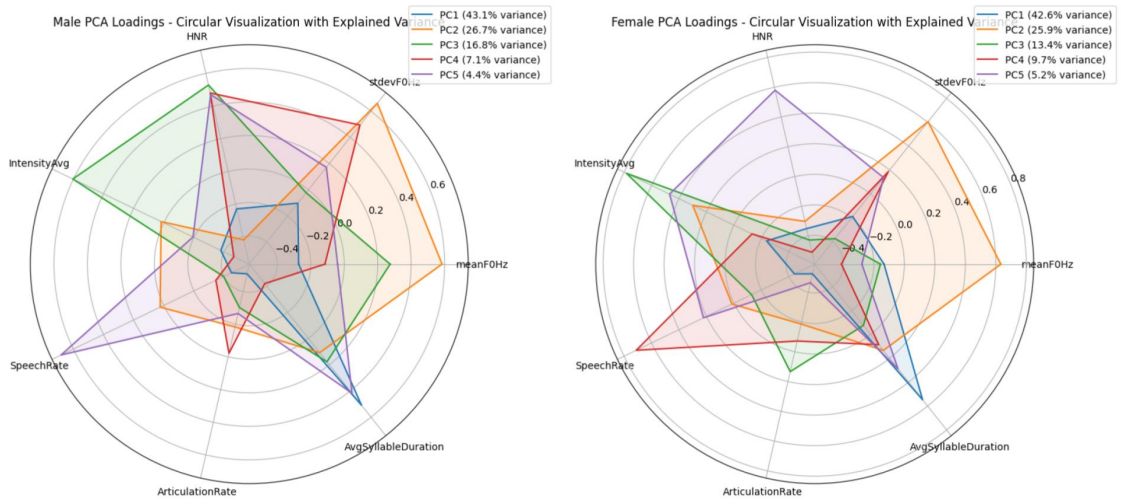


Figure 5.3: Circular visualization of PCA loadings for Kind Irony vs. Sincere (Anger): feature contributions to principal components for male (left) and female (right) speakers.

despite loading prominently on PC1. This apparent dissociation between PCA salience and inferential non-significance is theoretically informative: temporal features appear to function as a structural scaffold, preserving conversational fluency and maintaining the rhythmic baseline of ordinary interaction, while affective modulation is carried exclusively through spectral and energetic cues.

Sarcasm vs Sincere (Happy)

The prosodic profile of sarcasm stands in systematic contrast to that of kind irony. Rather than attenuating affective expression, sarcastic speech amplifies multiple acoustic dimensions simultaneously, producing a pattern of marked vocal intensification relative to the sincere happy baseline. This amplification strategy is consistent across genders, though the temporal dimension exhibits a notable gender asymmetry.

Pitch and Intensity as Primary Markers of Sarcasm. Both male and female speakers produce significantly elevated mean fundamental frequency in sarcastic speech (males: $+3.26$ semitones, $p < .05$; females: $+5.10$ semitones, $p < .05$), alongside substantial increases in vocal intensity (males: $+8.21$ dB, females: $+9.85$ dB, $p < .01$ for both). These effects position pitch elevation and loudness enhancement as the primary acoustic correlates of sarcasm, a finding reflected in the PCA structure: for both genders, MeanF0 and IntensityAvg load prominently on PC1, which accounts for over 41% of variance for males and approaches 50% for females.

Unlike kind irony, in which pitch lowering occurred without changes in variability, sarcasm exhibits a tendency toward increased pitch dispersion, particularly among female speakers. Elevated StdvF0 values and strong loadings on PC2 and PC3 suggest a more dynamically modulated

Gender	Feature	Sincere		Sarcasm		Test	P-value	Sig.
		Mean	Stdv	Mean	Stdv			
Female	meanF0Hz	192.38	20.32	258.33	47.98	Welch t-Test	0.018521	True
	stdevF0Hz	45.38	21.01	66.89	24.18	Student t-test	0.059609	False
	HNR	8.86	2.24	9.36	2.51	Student t-test	0.664115	False
	IntyAvg	53.36	4.39	63.21	9.12	Welch t-Test	0.044945	True
	SpeechRte	3.22	1.08	3.31	1.00	Student t-test	0.870784	False
	ArtRte	3.69	0.90	4.54	0.73	Student t-test	0.055378	False
	AvgSylDur	0.29	0.08	0.23	0.04	Mann-Whitney U	0.109340	False
Male	meanF0Hz	139.85	33.23	168.84	31.23	Student t-test	0.012149	True
	stdevF0Hz	49.29	25.23	48.58	19.91	Mann-Whitney U	1.000000	False
	HNR	8.26	1.73	9.33	1.62	Student t-test	0.069803	False
	IntAvg	57.40	8.39	65.61	5.18	Mann-Whitney U	0.002413	True
	SpeechRte	2.98	0.89	2.85	0.95	Student t-test	0.670947	False
	ArtRte	3.15	0.79	4.21	0.95	Mann-Whitney U	0.001740	True
	AvgSylDur	0.34	0.08	0.26	0.12	Mann-Whitney U	0.001740	True

Table 5.4: Statistical Analysis of Vocal Features by Gender and Expression Type (Sincere vs. Sarcasm)

intonational profile in sarcastic delivery, one characterized by wider pitch excursions and more pronounced contour variation.

Temporal Acceleration and Gender Asymmetry. Temporal features play a more substantive role in sarcasm than in kind irony, though their contribution is gender-differentiated. Male speakers show significantly higher articulation rate and shorter average syllable duration in sarcastic conditions ($p < .01$ for both), indicating increased speech motor activity and temporal compression. These effects are absent in female speakers, for whom no temporal feature reaches significance. In the PCA space, SpeechRate and ArticRate load substantially on higher-order components (PC3–PC5) for male speakers, suggesting that temporal densification constitutes a salient secondary dimension of male sarcastic expression, operating in conjunction with pitch and intensity amplification to produce an acoustically maximally marked signal. The gender asymmetry in temporal behavior implies that while both male and female sarcasm are grounded in energetic intensification, male speakers additionally recruit acceleration of articulation as an expressive resource.

Voice Quality and Secondary Acoustic Dimensions. Harmonics-to-noise ratio does not differ significantly between conditions for either gender and is distributed across higher-order PCA components, indicating that phonatory quality plays a subordinate role in sarcasm. This distinguishes sarcasm from kind irony, where female speakers exhibited a cleaner phonatory signal as part of the ironic configuration. Sarcasm, in contrast, does not appear to require systematic regulation of phonatory periodicity, relying instead on amplitude, pitch, and, for male speakers, temporal compression to encode its pragmatic function.

Taken together, the PCA configurations for sarcasm reveal a prosodic structure organized around amplification and contrast. The dominance of pitch and intensity in PC1, combined with

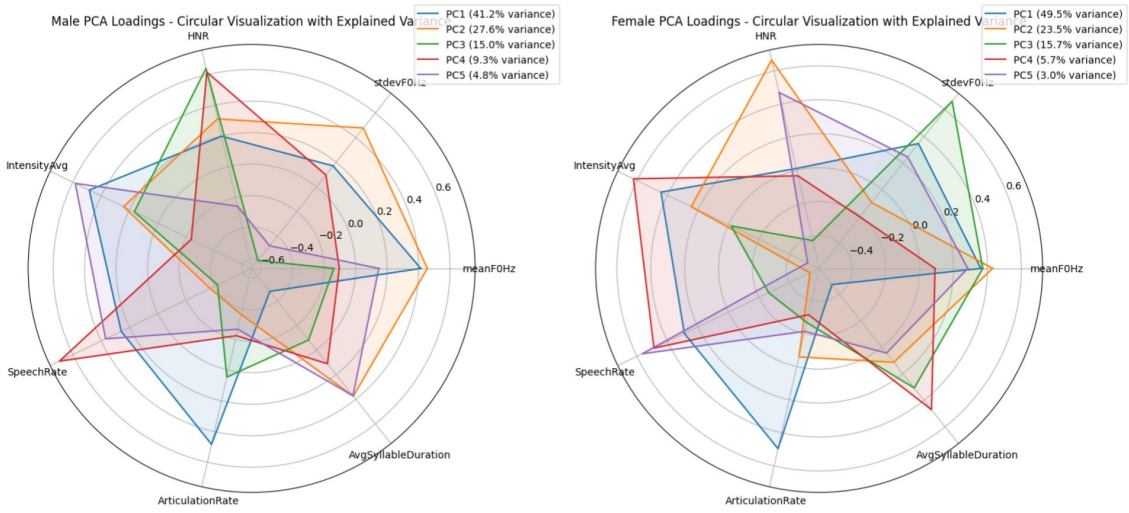


Figure 5.4: Circular visualization of PCA loadings for Sarcasm vs. Sincere (Happy): feature contributions to principal components for male (left) and female (right) speakers.

the recruitment of temporal features in subsequent components for male speakers, suggests that sarcasm is produced by systematically exceeding the prosodic norms associated with sincere happy speech, generating an acoustically anomalous signal whose implausibility relative to the emotional baseline constitutes the primary perceptual cue to non-literal intent.

5.3 Perceptual Evaluation

The acoustic analyses reported above establish that kind irony and sarcasm are associated with distinct and internally coherent prosodic configurations that systematically differ from their sincere counterparts along multiple acoustic dimensions. A remaining and non-trivial question concerns whether these configurations are perceptually recoverable, that is, whether the prosodic signatures identified through statistical inference are sufficient to support human listeners' judgments of ironic intent in the absence of contextual or interactional scaffolding. This question was addressed through a controlled perceptual evaluation.

Fifty-two audio stimuli drawn from the TIM corpus were evaluated by six native speakers of American English. All stimuli were produced in American English and distributed across four pragmatic-affective categories: sarcastic speech, kind-ironic speech, sincere happy speech, and sincere anger speech. Each stimulus consisted of a single utterance presented in isolation, and listeners were asked to rate the degree to which the speaker's intent was perceived as ironic using a five-point Likert scale. Stimuli were evaluated independently in a single-audio listening paradigm administered via the Podonos platform. The full evaluation yielded 312 ratings (52 stimuli $\times$ 6 evaluators). To avoid pseudo-replication, ratings were aggregated at the stimulus level prior to analysis, treating each stimulus as an independent observational unit.

Category	Mean	Median	Std	SEM	95% CI
Kind-ironic	3.308	3.538	1.027	0.459	1.181
Sarcastic	3.795	3.962	1.034	0.462	1.189
Sincere anger speech	2.064	1.615	1.219	0.545	1.402
Sincere happy speech	2.449	2.385	0.960	0.429	1.104

Table 5.5: Descriptive statistics of perceived irony ratings (stimulus-level means) by category.

Given the ordinal measurement scale and the infeasibility of assuming normality for stimulus-level distributions, non-parametric Mann-Whitney U tests were applied to two pre-specified contrasts. Effect sizes were quantified using Cliff's δ, with confidence intervals estimated via bootstrap resampling (10,000 iterations). Holm-Bonferroni correction was applied to control the family wise error rate across comparisons.

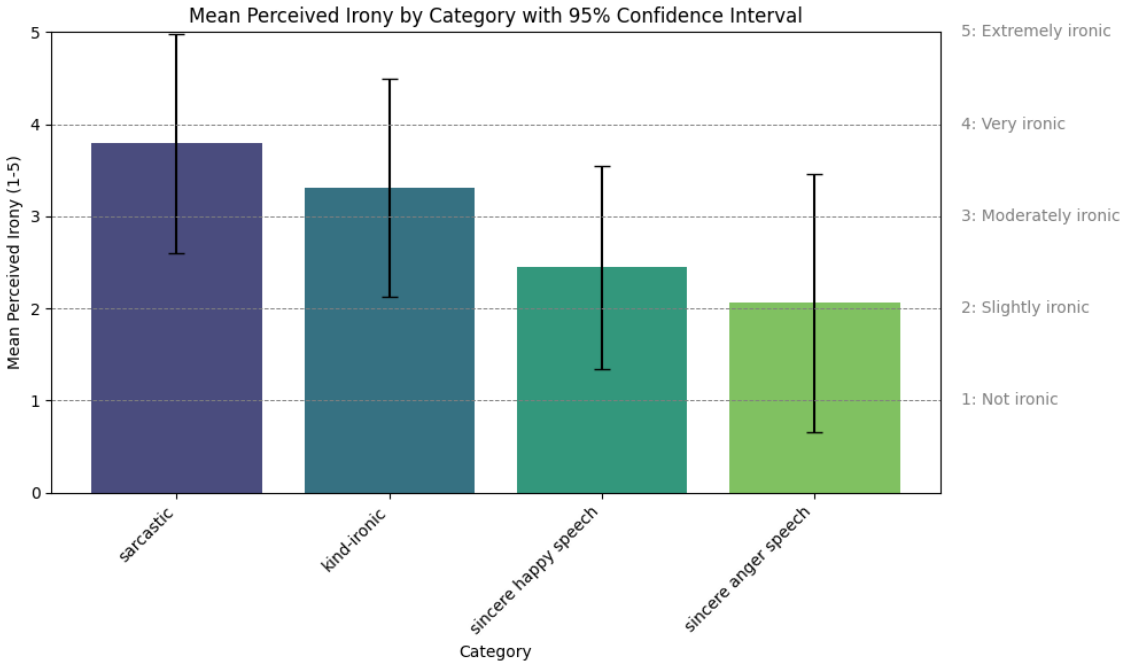


Figure 5.5: Mean irony ratings across affective speech categories. Error bars represent 95% confidence intervals.

Kind-Ironic vs. Sincere (Anger)

Kind-ironic stimuli ($n = 13$) received substantially higher perceived irony ratings than sincere anger stimuli ($n = 13$), with stimulus-level means of 3.31 and 2.06, respectively. This difference was statistically significant ($U = 149.0$, $p = .0010$) and survived Holm correction ($p_{\text{adj}} = .0019$). The effect was large (Cliff's $\delta = 0.76$, 95% CI $[0.46, 0.99]$), corresponding to a 76% probability that a randomly selected kind-ironic stimulus would be rated as more ironic than a randomly selected sincere anger stimulus.

Sarcastic vs. Sincere (Happy)

Sarcastic stimuli ($n = 13$) similarly elicited higher perceived irony than sincere happy stimuli ($n = 13$), with means of 3.80 and 2.45, respectively. The Mann–Whitney U test yielded a significant result ($U = 146.5$, $p = .0016$, $p_{\text{adj}} = .0019$), and the effect was again large (Cliff's $\delta = 0.73$, 95% CI $[0.40, 0.98]$), corresponding to a 73% probability that a sarcastic stimulus would be rated more ironic than a sincere happy stimulus. Considered jointly, the two contrasts yield a coherent and convergent pattern. Both irony types are judged as significantly more ironic than their sincere counterparts, with large, robust effect sizes that survive multiple comparison correction.

Summary and Discussion

This chapter characterizes dripping prosodic irony as a systematic *prosodic–semantic divergence*, showing that two pragmatic subtypes—kind irony and sarcasm—are realized through opposing acoustic strategies: **attenuation** (pitch ↓, intensity ↓) and **amplification** (pitch ↑, intensity ↑), respectively. These patterns are robust, statistically significant, and perceptually recoverable, supporting the view that prosody encodes not only non-literal meaning but also *affective stance*.

In line with Anolli, Ciceri, and Infantino (2002), our results confirm that irony is not defined by fixed acoustic markers but by *deviation from a contextual baseline*. Apparent contradictions with prior findings, such as increased vs. decreased pitch, are resolved when baseline selection is made explicit. For example, our attenuation in kind irony aligns with Scharrer, Christmann, and Knoll (2011) (lower F0), while differences in energy measures reflect methodological scope. Similarly, the coexistence of pitch lowering (this study) and pitch range expansion (B. Braun and Anna Schmiedel (2018)) suggests that mean level and variability operate as partially independent dimensions.

For sarcasm, our findings of prosodic amplification (pitch and intensity increase) converge with Italian and American data (Anolli, Ciceri, and Infantino 2002; Rockwell 2000), but contrast with the "deadpan" flattening reported in German (Cheang and Pell 2008; B. Braun and Anna Schmiedel 2018). This divergence is best understood not as inconsistency but as evidence for language-specific realization strategies, where speakers exploit different prosodic norms to signal markedness. In this sense, we support Fónagy (1971)'s early claim that irony is inherently multi-dimensional, while extending it with a cross-linguistic perspective: irony is universally *relational*, but locally instantiated.

The opposition between attenuation (kind irony) and amplification (sarcasm) also aligns with pragmatic theories distinguishing affiliative vs. critical irony (Kreuz and Glucksberg 1998; Gibbs 2000). Our data demonstrate that this distinction is not merely semantic but acoustically encoded: restrained prosody mitigates negative content in kind irony, while exaggerated prosody undermines positive content in sarcasm. This supports a unified account in which prosody functions as a carrier of speaker stance, modulating the interpretation of literal meaning.

Perceptually, the significant effect sizes observed (Cliff's $\delta \approx 0.73$–0.76) confirm that prosody alone provides sufficient cues for irony detection, consistent with Bryant and Fox Tree (2005). However, moderate absolute ratings reinforce Mehrabian and Wiener (1967)'s view that meaning emerges from multimodal integration: prosody is powerful but not exhaustive.

Finally, gender-differentiated secondary features—temporal acceleration in male sarcasm and increased voice quality in female kind irony, extend observations by B. Braun and Anna Schmiedel (2018) and align with sociolinguistic accounts of performative identity (Brown and Levinson 1987). These results suggest that while core acoustic dimensions are shared, their configuration reflects social norms and communicative style.

Chapter 6

Affective Virtual Environment Generation

Understanding how specific audiovisual features contribute to emotional expression is critical for designing affective virtual environments. In this chapter, we present strategies for translating statistically significant audiovisual features into structured prompts for affective environment generation. Our aim is to evaluate whether descriptors derived from empirical analyses of audio and visual datasets can meaningfully inform prompt design. By identifying acoustic and visual features that reliably differentiate emotional categories, we construct semantic profiles that serve as intermediate representations for guiding generative models.

The chapter is organized into three main sections. First, Section 6.1 analyzes musical features extracted from the 4Q Audio Emotion Database to identify acoustic attributes that drive emotional differentiation across Russell's circumplex quadrants. This analysis focuses on how dimensions such as harmony, timbre, loudness, and rhythmic activity relate to perceived emotional states in music. Next, Section 6.2 examines visual features from the Emotion6 dataset to determine how graphical properties contribute to emotional categorization. By analyzing color distributions, textural descriptors, structural complexity, and luminance characteristics, we highlight the visual cues that most consistently differentiate emotional categories.

Finally, Section 6.3 integrates findings from both analyses into a three-level prompt engineering framework for affective generation. This framework translates statistically significant audiovisual descriptors into prompts with increasing levels of specificity, ranging from simple emotional labels to feature-based semantic descriptions and approximate parametric constraints. These prompts are used to experimentally assess whether empirically derived descriptors can influence the generation of affective audiovisual environments, while also revealing the practical limits of prompt-based control in contemporary generative models.

For transparency and reproducibility, the full implementation of the experimental pipelines described in this chapter is publicly available. The corresponding Colab notebooks can be accessed here:

- Affective Music Analysis Notebook
- Affective Visual Analysis Notebook

6.1 Affective Music Generation

This section reports the results of the statistical feature analysis conducted on the 4Q Audio Emotion Database (Panda, Malheiro, and Paiva 2018), following the inferential pipeline described in Section 3.3.1. The analysis aims to identify the acoustic attributes that most reliably differentiate emotional categories, providing an empirically grounded reference for composing semantic prompts in text-guided music generation systems.

6.1.1 Data and Feature Selection

The dataset comprises Pop/Rock excerpts filtered from the 4Q corpus (see Figure 6.1), selected on the basis of their dominant representation and affective coherence within the collection. Within this subset, clips were identified using emotional tags: *happy* (21), *sad* (20), *anger* (58), and *gentle* (22), each corresponding to one quadrant of Russell's circumplex model (see Figure 6.2). The dataset was balanced through random downsampling to 20 samples per category prior to analysis. Features were extracted using the Essentia library across five musical dimensions: melody, harmony, rhythm, dynamics, and timbre.

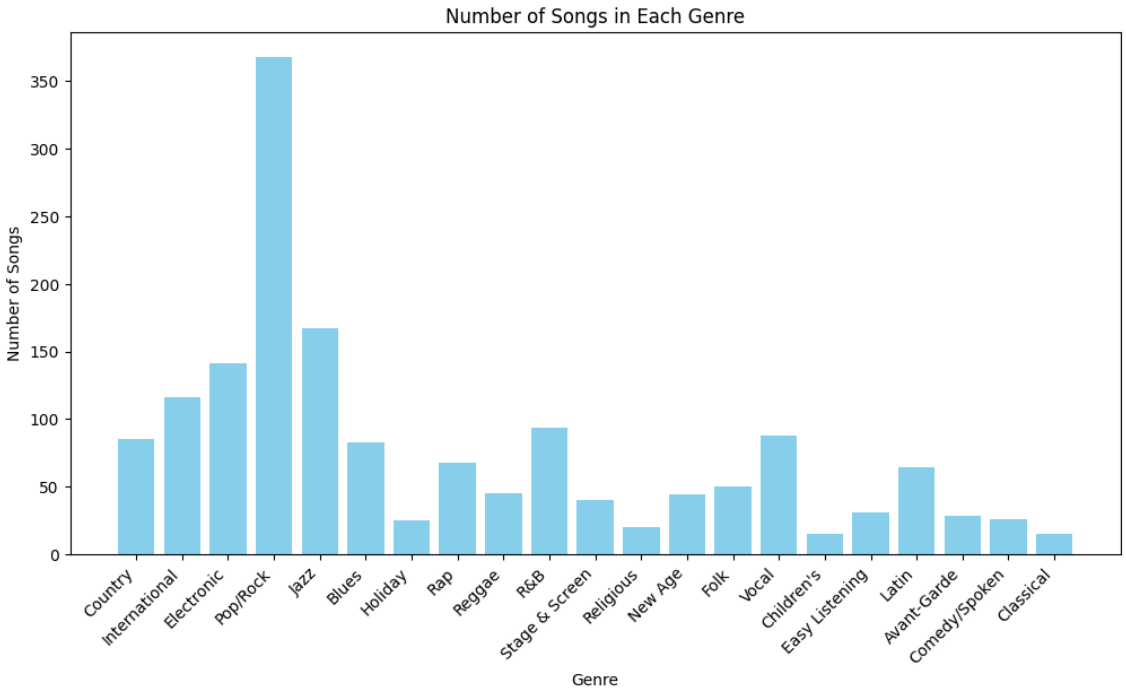


Figure 6.1: Distribution of musical excerpts across genres in the 4Q Audio Emotion Database. Pop/Rock represents the largest category, followed by Jazz and Electronic music. This distribution motivated the selection of the Pop/Rock subset for the affective feature analysis.

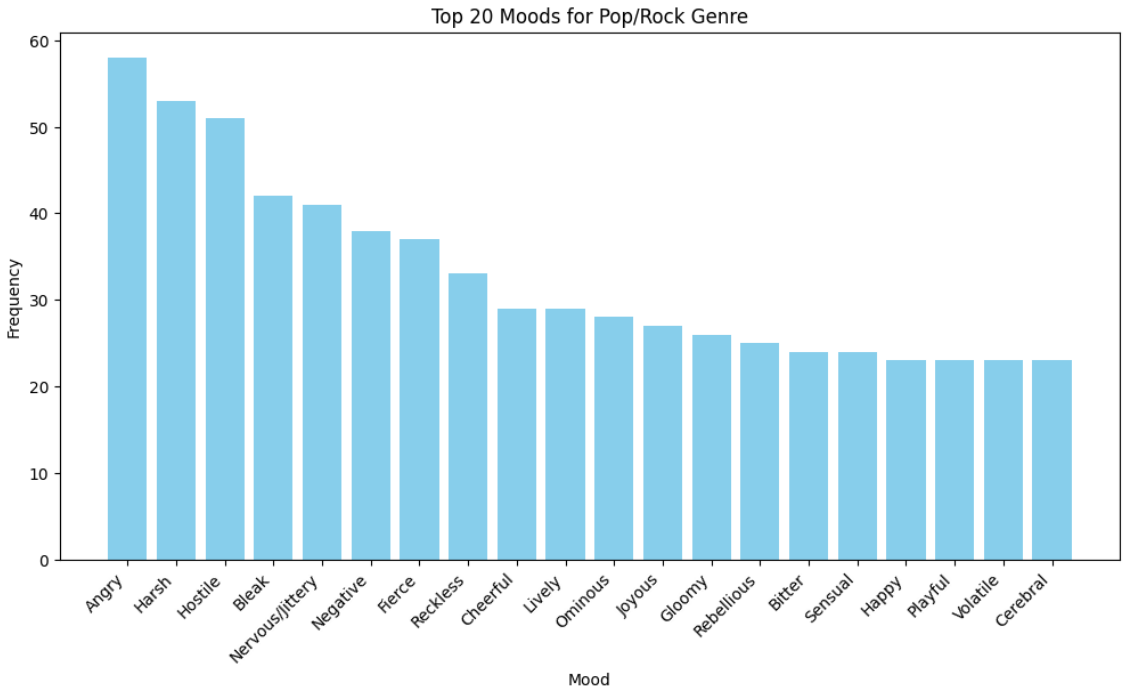


Figure 6.2: Frequency of top mood tags in Pop/Rock tracks. High-arousal negative descriptors (*angry*, *harsh*, *hostile*) co-occur with positive descriptors (*happy*, *joyous*), confirming the suitability of this subset for valence-arousal analysis.

6.1.2 Statistical Analysis

Across all musical dimensions examined, 11 of the 17 extracted features surpassed Benjamini–Hochberg FDR correction ($\alpha = 0.05$), collectively constituting an empirically grounded acoustic signature for each emotional quadrant. Effect sizes ranged from moderate ($\eta^2 = 0.092$ for pitch salience standard deviation, non-significant after correction) to very large ($\eta^2 = 0.791$ for HPCP entropy mean), indicating that harmonic descriptors carry substantially more discriminative information than rhythmic or melodic features in this corpus. Descriptive statistics for all features are reported in Table 6.1, with FDR-surviving features marked by an asterisk.

Consistent with an arousal–valence interpretation of Russell's circumplex model, the discriminative structure of the feature space partitions primarily along the *arousal* axis: high-arousal categories (*Angry*, *Happy*) cluster toward elevated onset density, loudness, pitch salience, and spectral brightness, while low-arousal categories (*Sad*, *Tender*) exhibit the inverse pattern across the same dimensions. Tonal mode and harmonic complexity provide the principal gradient along the *valence* axis. The distributional spread of discriminative features across quadrants is visualized in Figure 6.3.

Harmonic Dimensions

Harmonic descriptors produced the largest and most consistent effect sizes of any dimension, with HPCP entropy mean constituting the single most discriminative feature in the entire analysis

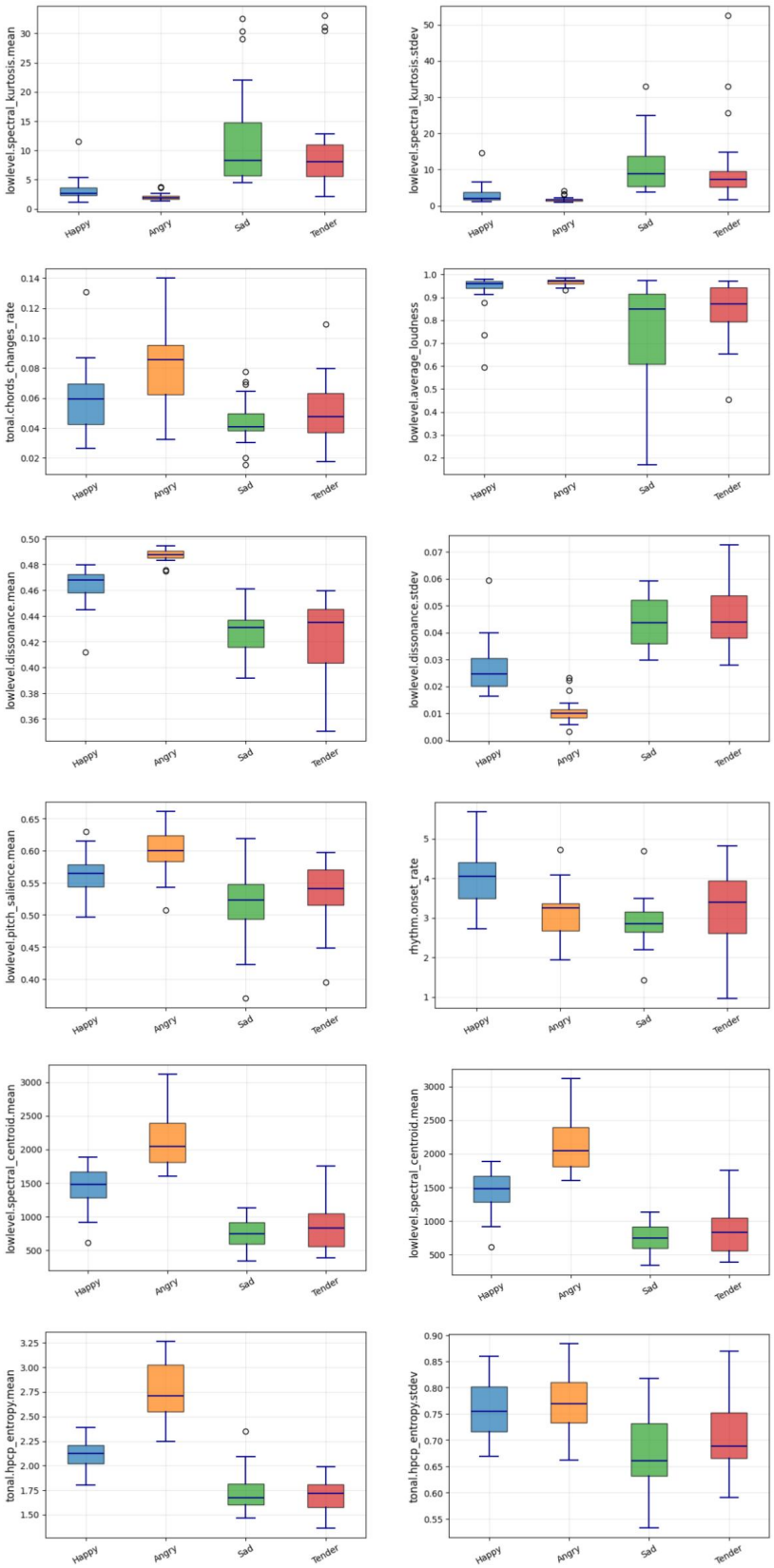


Figure 6.3: Box plots showing the distribution of discriminative musical features across the four emotional quadrants. Variations in spectral, harmonic, rhythmic, and dynamic descriptors confirm differential affective profiles and are consistent with the arousal–valence structure of Russell's circumplex model. Features marked with an asterisk (*) in Table 6.1 survived FDR correction.

Feature	Units	Happy (Q1)	Angry (Q2)	Sad (Q3)	Tender (Q4)
rhythm.bpm	BPM	129.35 ± 21.26	122.03 ± 23.94	117.79 ± 26.66	123.13 ± 24.47
spectral_kurtosis.mean*	unitless	3.29 ± 2.15	2.12 ± 0.76	12.32 ± 8.86	11.19 ± 8.93
spectral_kurtosis.stdev*	unitless	3.20 ± 2.91	1.83 ± 0.75	11.47 ± 7.82	11.34 ± 11.98
chords_number_rate	prop.(0–1)	0.013 ± 0.004	0.012 ± 0.004	0.013 ± 0.004	0.012 ± 0.004
chords_changes_rate*	prop. (0–1)	0.061 ± 0.024	0.085 ± 0.027	0.045 ± 0.016	0.050 ± 0.021
dissonance.mean*	0–1	0.463 ± 0.016	0.487 ± 0.005	0.427 ± 0.018	0.424 ± 0.028
dissonance.stdev*	0–1	0.027 ± 0.010	0.011 ± 0.005	0.044 ± 0.009	0.046 ± 0.012
average_loudness*	0–1	0.925 ± 0.093	0.968 ± 0.013	0.747 ± 0.227	0.834 ± 0.135
pitch_salience.mean*	0–1	0.561 ± 0.033	0.599 ± 0.037	0.514 ± 0.055	0.532 ± 0.047
pitch_salience.stdev	0–1	0.107 ± 0.014	0.103 ± 0.017	0.119 ± 0.024	0.111 ± 0.017
rhythm.onset_rate*	onsets/s	4.076 ± 0.737	3.164 ± 0.643	2.878 ± 0.616	3.312 ± 0.929
spectral_centroid.mean*	Hz	1432.68 ± 346.15	2141.69 ± 416.99	758.88 ± 194.14	844.92 ± 343.76
spectral_centroid.stdev	Hz	667.24 ± 149.82	567.67 ± 178.29	579.87 ± 139.95	599.84 ± 137.24
hpcp_entropy.mean*	entropy	2.098 ± 0.158	2.746 ± 0.298	1.742 ± 0.220	1.697 ± 0.156
hpcp_entropy.stdev*	entropy	0.760 ± 0.053	0.769 ± 0.054	0.677 ± 0.070	0.703 ± 0.067
key_krumhansl.scale*	maj/min	80% / 20%	40% / 60%	60% / 40%	90% / 10%
chords_key	categ.	F, C, A, D, E	C#, C, G, A	E, C, G, Ab	F, G, Bb, E

Table 6.1: Comparison of musical features across emotional quadrants (mean $\pm$ standard deviation). Asterisks (*) denote features surviving Benjamini–Hochberg FDR correction ($\alpha = 0.05$). Non-significant features are included for completeness.

($\eta^2 = 0.791$, $p < 0.001$). *Angry* excerpts displayed markedly more complex and unpredictable harmonic profiles (2.746 ± 0.298) relative to *Tender* and *Sad* categories, which converged on low-entropy, tonally stable configurations (1.697 ± 0.156 and 1.742 ± 0.220, respectively).

Dissonance mean and standard deviation yielded comparably large effects ($\eta^2 = 0.676$ and $\eta^2 = 0.689$, $p < 0.001$ for both). Crucially, these two measures capture qualitatively distinct acoustic phenomena: *Angry* clips are characterized by the highest mean dissonance (0.487 ± 0.005) combined with the *lowest* variability (0.011 ± 0.005), indicating sustained harmonic tension with minimal dynamic modulation. Conversely, *Tender* and *Sad* excerpts exhibit lower mean dissonance but substantially higher standard deviation (0.046 ± 0.012 and 0.044 ± 0.009).

Chord change rate ($\eta^2 = 0.323$, $p < 0.001$) further differentiates along the arousal axis: *Angry* clips sustain the fastest harmonic rhythm (0.085 ± 0.027), while *Sad* excerpts show the most static harmonic progressions (0.045 ± 0.016).

Tonal mode distribution confirmed a significant association with emotional quadrant ($\chi^2 = 13.447$, $p = 0.004$): major modes predominated in *Happy* (80%) and *Tender* (90%) excerpts, while *Angry* clips favored minor tonalities (60%). Notably, *Sad* clips deviated from the conventional valence–mode mapping, exhibiting a mixed distribution (60% major, 40% minor).

Timbral Dimension

Spectral centroid mean ranked second overall in discriminative power ($\eta^2 = 0.732$, $p < 0.001$), confirming timbral brightness as a robust arousal marker. *Angry* clips were nearly three times brighter than *Sad* excerpts ($2{,}141.69 \pm 416.99$ Hz vs. 758.88 ± 194.14 Hz), with *Happy* at an intermediate level ($1{,}432.68 \pm 346.15$ Hz). Spectral kurtosis, both mean ($\eta^2 = 0.338$) and standard deviation ($\eta^2 = 0.272$, $p < 0.001$ for both), inverted this pattern: *Sad* and *Tender* clips exhibited substantially higher kurtosis values (12.32 ± 8.86 and 11.19 ± 8.93, respectively), indicating

narrower, more peaked spectral distributions corresponding to a more resonant, sustained timbral character associated with low-arousal sound production. Spectral centroid standard deviation did not survive correction ($p = 0.068$, $\eta^2 = 0.060$), indicating that timbral variability, as opposed to mean brightness, carries limited affective signal in this corpus.

Melodic and Dynamic Dimensions

Pitch salience mean significantly discriminated quadrants ($\eta^2 = 0.347$, $p < 0.001$), with *Angry* clips exhibiting the sharpest pitch definition (0.599 ± 0.037)—consistent with the focused vocal effort and narrower melodic register typical of high-arousal musical expression—and *Sad* excerpts showing the greatest tonal ambiguity (0.514 ± 0.055). Onset rate ($\eta^2 = 0.263$, $p < 0.001$) was the strongest rhythmic predictor, with *Happy* clips showing the highest attack density (4.076 ± 0.737 onsets/s), significantly exceeding all other quadrants. The failure of global tempo (BPM) to differentiate quadrants ($\eta^2 = 0.029$, $p = 0.340$) despite all categories falling within a narrow range ($\approx$ 118–129 BPM) supports the conclusion that rhythmic density, rather than tempo per itself, mediates perceived emotional energy in this genre. Average loudness ($\eta^2 = 0.270$, $p < 0.001$) followed the arousal gradient predictably: *Angry* clips were the loudest (0.968 ± 0.013), while *Sad* clips were the quietest (0.747 ± 0.227).

Section Summary

Emotional differentiation in music emerges from a constellation of features rather than any single descriptor. High-arousal categories (*Angry*, *Happy*) are jointly characterized by elevated onset density, louder dynamics, higher pitch salience, and brighter spectral profiles. Low-arousal categories (*Sad*, *Tender*) present the inverse pattern. Harmonic descriptors—particularly HPCP entropy ($\eta^2 = 0.791$), dissonance ($\eta^2 = 0.676$–0.689), and chord change rate ($\eta^2 = 0.323$)—consistently produced the largest effect sizes, underscoring the centrality of tonal instability and harmonic tension in affective perception. Notably, global tempo failed to differentiate categories in this dataset, suggesting that rhythmic density and timbral energy are more functionally relevant to emotional characterization.

6.2 Affective Visual Generation

This section reports the results of the statistical feature analysis conducted on the Emotion6 database, following the inferential pipeline described in Section 3.3.2. The analysis aims to identify the low-level visual properties that most reliably differentiate emotional categories, providing an empirically grounded reference for composing semantic prompts in text-guided image generation systems.

6.2.1 Data and Feature Selection

The dataset comprises naturalistic images from the *Emotion6 (ROI)* corpus (Peng et al. 2015), sourced from Flickr and equally balanced across six basic emotions: *anger*, *disgust*, *fear*, *joy*, *sadness*, and *surprise* (330 images each). Annotations were collected via Amazon Mechanical Turk with 15 independent raters per image. No down-sampling was required given the balanced class distribution. Features were extracted across five visual dimensions: color, texture, image complexity, symmetry, and structural composition.

6.2.2 Statistical Analysis

Across all visual dimensions examined, 54 of the 726 extracted features survived Benjamini–Hochberg FDR correction ($\alpha = 0.05$), collectively constituting an empirically grounded visual signature for each emotional category. Effect sizes ranged from small ($\eta^2 \approx 0.01$) to medium magnitudes ($\eta^2 \approx 0.11$), indicating that color and texture descriptors carry substantially more discriminative information than structural geometry features. The Joy–Anger pairwise analysis confirmed and extended these findings: Mann–Whitney U tests identified robust differences across 464 features, reinforcing that the valence contrast between these two emotions is strongly expressed in low-level chromatic and textural properties. Descriptive statistics for the most discriminative features across emotional categories are reported in Table 6.2, with FDR-surviving features marked by an asterisk.

Feature	Anger	Disgust	Fear	Joy	Sadness	Surprise
Brightness*	107.6 ± 43.4	111.3 ± 35.8	81.1 ± 40.0	112.1 ± 38.9	106.4 ± 44.3	123.6 ± 38.3
Entropy*	6.93 ± 0.97	7.22 ± 0.61	6.66 ± 1.22	7.21 ± 0.57	6.79 ± 1.09	7.11 ± 0.70
Edge Density*	20.99 ± 17.28	33.27 ± 19.80	25.05 ± 20.07	26.95 ± 17.18	21.79 ± 16.89	26.90 ± 19.20
Contrast*	56.22 ± 17.51	53.38 ± 15.01	57.18 ± 18.06	55.61 ± 16.34	58.72 ± 18.92	53.74 ± 15.48
Hue (mean)*	50.4 ± 37.1	45.5 ± 27.7	43.1 ± 35.0	56.3 ± 32.1	41.2 ± 36.5	63.9 ± 30.6
Saturation (mean)*	79.8 ± 60.6	89.2 ± 46.4	79.7 ± 61.2	109.3 ± 56.3	60.0 ± 51.5	98.8 ± 52.1
Gabor ch.0*	222.5 ± 38.2	243.0 ± 23.6	209.5 ± 53.8	239.3 ± 29.1	221.4 ± 41.7	231.6 ± 33.5
LBP_2*	0.046 ± 0.013	0.053 ± 0.013	0.045 ± 0.014	0.047 ± 0.013	0.040 ± 0.015	0.047 ± 0.013
H. Symmetry*	120.5 ± 13.2	124.7 ± 6.8	118.1 ± 17.6	122.3 ± 10.4	121.4 ± 11.9	121.8 ± 11.3
V. Symmetry*	121.0 ± 13.5	125.5 ± 6.1	118.6 ± 18.5	122.9 ± 10.6	121.8 ± 12.1	122.3 ± 11.2

Table 6.2: Comparison of selected visual features across emotional categories (mean $\pm$ standard deviation). Asterisks (*) denote features surviving Benjamini–Hochberg FDR correction ($\alpha = 0.05$). Values for brightness are in pixel intensity units (0–255); hue and saturation are in HSV scale (0–255); all other features are unitless proportions or normalized scores.

Color Dimensions

Color-based descriptors produced the largest and most consistent effect sizes of any dimension. Color moment features yielded the single most discriminative feature in the entire analysis ($\eta^2 = 0.109$, $p < 0.001$), with *Joy* exhibiting the highest brightness values (136.71 ± 43.22) and *Fear* the lowest (93.91 ± 43.70).

Saturation and hue distributions showed comparably strong effects across all nine color moment features (η^2: 0.009–0.109, $p < 0.01$ for all). These two measures capture qualitatively distinct chromatic phenomena: *Joy* is characterized by the highest saturation (109.3 ± 56.3) combined with warm hue values (56.3 ± 32.1), indicating vivid, energetically positive chromatic content. Conversely, *Sadness* exhibits the lowest saturation (60.0 ± 51.5) and the coldest hue distribution (41.2 ± 36.5), reflecting muted, desaturated palettes consistent with low-arousal negative affect. *Anger* occupies an intermediate chromatic position with high saturation variability (79.8 ± 60.6), indicating heterogeneous color usage across instances rather than a consistent warm or cool signature.

All nine dominant color features also showed significant differentiation (η^2: 0.005–0.038, $p < 0.05$ for all), with *Joy* consistently exhibiting the highest dominant color values and *Fear* the lowest, confirming that the vibrancy of the overall color palette tracks emotional valence. Ten of 512 color histogram bins survived FDR correction (η^2 up to 0.094), with the first histogram bin showing the strongest localized effect: *Fear* and *Sadness* dominated dark HSV regions, while *Joy* and *Disgust* showed the lowest activation, consistent with their brighter visual profiles (see Figure 6.4).

Textural Dimensions

Gabor filter responses ranked second overall in discriminative power (η^2: 0.071–0.093, $p < 0.001$ for all four orientations), confirming that directional textural energy is a robust affective marker analogous to spectral brightness in the musical domain. *Disgust* and *Joy* exhibited the highest Gabor responses (243.03 ± 23.55 and 239.27 ± 29.08 in channel 0 respectively), indicating more structured and patterned visual content. *Fear* showed the lowest responses (209.45 ± 53.84), suggesting less organized textural information in fear-evoking images. The consistency of these effects across all four Gabor orientations (0°, 45°, 90°, 135°) confirms that directional texture information varies systematically with emotional content.

Local Binary Pattern (LBP) descriptors provided complementary micro-textural information, with ten of 58 features surviving FDR correction (η^2: 0.020–0.080). The second LBP feature showed the strongest effect ($\eta^2 = 0.080$), with *Disgust* displaying the highest values (0.0529 ± 0.0126) and *Sadness* the lowest (0.0399 ± 0.0154). The eighth LBP feature further highlighted that *Anger* and *Fear* carry the most irregular micro-textural patterns (0.1937 ± 0.1481 and 0.2057 ± 0.1591 respectively), while *Disgust* showed the most regular surface structure (0.1227 ± 0.0705). HOG features showed weaker but significant discriminability, with six of 128 features surviving correction, confirming that coarse edge orientations contribute less than fine-grained textural encoding to emotional differentiation.

Structural and Complexity Dimensions

Visual entropy significantly discriminated emotional categories ($\eta^2 = 0.054$, $p < 0.001$), with *Disgust* and *Joy* displaying the highest information content (7.22 ± 0.61 and 7.21 ± 0.57 respectively),

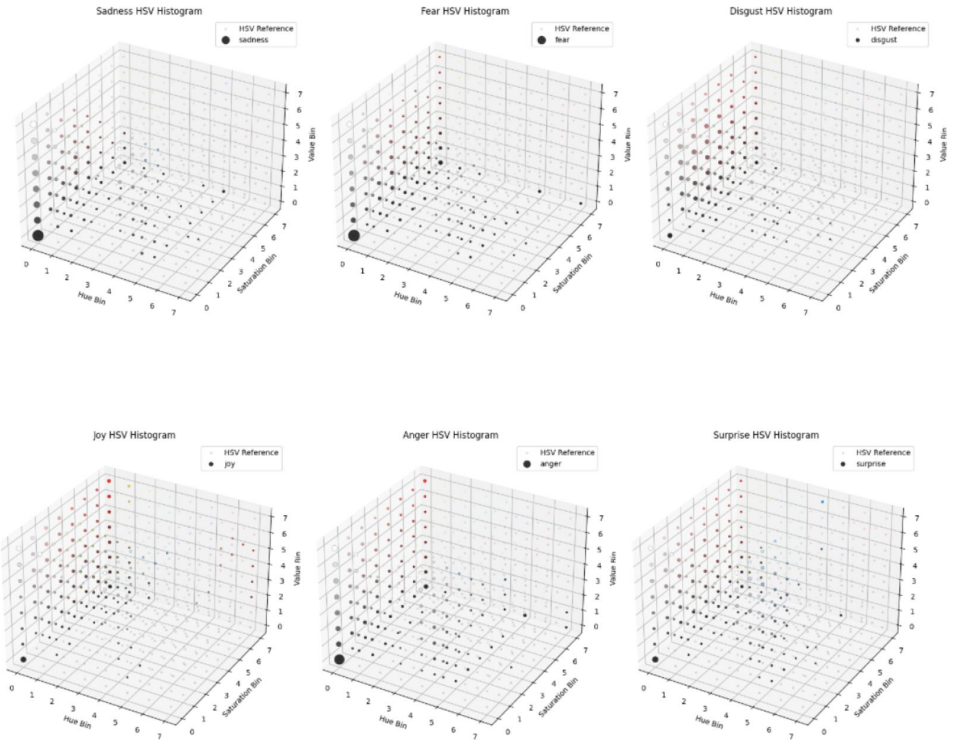


Figure 6.4: **3D Normalized HSV Color Histograms (512 Bins) across Emotional Categories.** Each plot visualizes the global color distribution using an $8 \times 8 \times 8$ binning scheme. The **spatial position and color** of each sphere correspond to its specific bin in the Hue-Saturation-Value coordinate system, while the **sphere radius** denotes the relative contribution (normalized frequency) of that bin to the emotional profile. Consistent with the statistical analysis, *Fear* and *Sadness* show pronounced activation (larger radii) in the low-Value and low-Saturation regions. Conversely, *Joy* exhibits a shift toward higher-Value bins and more vivid chromatic clusters, confirming its status as the most visually bright and saturated category ($\eta^2 = 0.109$).

and *Fear* and *Sadness* the lowest (6.66 ± 1.22 and 6.79 ± 1.09).

Brightness was the strongest structural predictor ($\eta^2 = 0.073$, $p < 0.001$), with *Joy* images the brightest (112.09 ± 38.91) and *Fear* the darkest (81.13 ± 40.03), confirming the well-established association between luminance and emotional valence. The failure of contrast to differentiate categories strongly ($\eta^2 = 0.013$, $p = 0.002$, though significant) suggests that luminance variability, as opposed to mean brightness, carries limited independent affective signal.

Edge density significantly distinguished emotional categories ($\eta^2 = 0.053$, $p < 0.001$), with *Disgust* exhibiting the highest structural complexity (33.27 ± 19.80) and *Anger* and *Sadness* the lowest (20.99 ± 17.28 and 21.79 ± 16.89). Both horizontal and vertical symmetry also showed significant differentiation ($\eta^2 = 0.041$ and 0.048 respectively, $p < 0.001$), with *Disgust* exhibiting the most balanced compositions and *Fear* the most asymmetric, indicating that compositional regularity contributes to affective interpretation independently of chromatic content.

Section Summary

Emotional differentiation in images emerges from a constellation of visual properties rather than any single descriptor. Positive categories (*Joy*, *Surprise*) are jointly characterized by elevated brightness, higher color saturation, stronger Gabor textural energy, and greater visual entropy. Negative categories (*Fear*, *Sadness*) present the inverse pattern across the same dimensions. Color descriptors, particularly color moments (η^2 up to 0.109), dominant color vectors (η^2 up to 0.038), and HSV histogram bins (η^2 up to 0.094), consistently produced the largest effect sizes, underscoring the centrality of chromatic intensity and luminance in affective visual perception. Notably, global contrast showed only modest discriminative power despite reaching significance, suggesting that mean luminance rather than luminance variability is the functionally relevant structural descriptor. Micro-textural and directional features (Gabor, LBP) provided robust complementary signal, while coarse structural geometry (HOG) contributed comparatively less to emotional differentiation.

6.3 A Three-Level Prompt Engineering Framework for Affective Environment Generation

The statistical results obtained from both the musical and visual analyses provide a framework for designing affective prompts that can guide generative AI models with varying levels of specificity.

We propose a three-level prompting framework that translates statistically significant features into conceptual, semantic, and parametric descriptors.

The framework consists of:

- **Level 1 – High-level emotional prompts** (conceptual; one emotional label)
- **Level 2 – Mid-level descriptive prompts** (semantic; feature-derived adjectives grounded in statistically significant effects)
- **Level 3 – Low-level parametric prompts** (numeric; approximate feature values extracted from the dataset)

These levels correspond to increasing model steering power: Level 1 describes *what* emotion to evoke; Level 2 describes *how* to evoke it; Level 3 describes *with what measurable properties* it should be generated.

To build ironic environments we focused on two opposite emotional categories: *Joy/Happy* (positive valence and *Anger* (negative valence). Middle and low level profiles were derived strictly from features that showed statistically significant effects after multiple-comparison correction, ensuring that the prompting strategy is grounded in empirically validated emotional cues.

6.3.1 Perceptual Evaluation

Using the emotional profiles obtained from the musical and visual analyses, a set of immersive audiovisual environments was generated. Visual scenes were created as equirectangular textures for 360° rendering, allowing participants to explore the environment freely while maintaining a simplified graphical structure suitable for controlled experimentation. The audio component consisted of short 10-second generative music excerpts produced from the corresponding musical prompts.

The environments were generated using the system developed in this research and deployed on the Hugging Face platform through a Gradio interface[1]. The system integrates automatic speech emotion recognition, sentiment analysis, and generative models for both image and audio synthesis. Acoustic emotion recognition is performed using a neural network trained on the RAVDESS dataset, while speech transcription is carried out using the Whisper model. Sentiment polarity is derived from the transcribed text using TextBlob. Based on these predictions, prompts are generated for both visual and musical synthesis. Visual environments are produced through a text-to-image generation pipeline that creates abstract equirectangular textures, while the audio

1. https://huggingface.co/spaces/jfforero/Affective_Virtual_Environmets

Table 6.3: Three-Level Prompting Framework for Affective Environment Generation

Emotion	Modality	Level	Prompt Description
Joy / Happy	Music	High-Level	*"Generate a happy soundtrack."*
		Mid-Level	*"Generate a soundtrack with clear melody, fast onset density, major mode harmony, low dissonance, bright timbre, and moderate loudness."*
		Low-Level	*"Generate music around 4.0 onsets/sec, ~129 BPM, pitch salience* ≈ 0.56*, major key, dissonance* ≈ 0.46*, harmonic entropy* ≈ 2.10*, spectral centroid* ≈ 1430 *Hz, loudness* ≈ 0.92*."*
	Image	High-Level	*"Generate a joyful equirectangular abstract texture image"*
		Mid-Level	*"Generate a vivid warm hues, high brightness, strong saturation, balanced symmetry, smooth textures, high entropy, and fine distributed edges, equirectangular abstract texture image."*
		Low-Level	*"Generate a equirectangular abstract texture image with HSV brightness* ≈ 137*, saturation* ≈ 109*, entropy* ≈ 7.21*, edge density* ≈ 27*, high symmetry, and low Gabor energy."*
Anger	Music	High-Level	*"Generate an anger soundtrack."*
		Mid-Level	*"Generate music with sharp pitch clarity, rapid harmonic changes, high dissonance, complex chromaticity, bright and aggressive timbre, and strong loudness in a minor mode."*
		Low-Level	*"Generate music around 3.1 onsets/sec, ~122 BPM, pitch salience* ≈ 0.60*, minor key, dissonance* ≈ 0.49*, harmonic entropy* ≈ 2.75*, spectral centroid* ≈ 2140 *Hz, loudness* ≈ 0.97*."*
	Image	High-Level	*"Generate an angry scene with dark, red-dominant tones and strong contrast."*
		Mid-Level	*"Generate an image with dark red-dominant tones, strong contrast, reduced brightness, sharp oriented textures, high Gabor energy, pronounced asymmetry, and low entropy, equirectangular abstract texture image."*
		Low-Level	*"Generate an image with brightness* ≈ 108*, saturation* ≈ 80*, entropy* ≈ 6.66*, edge density* ≈ 21*, strong Gabor responses, low symmetry, and red-dominant hues."*

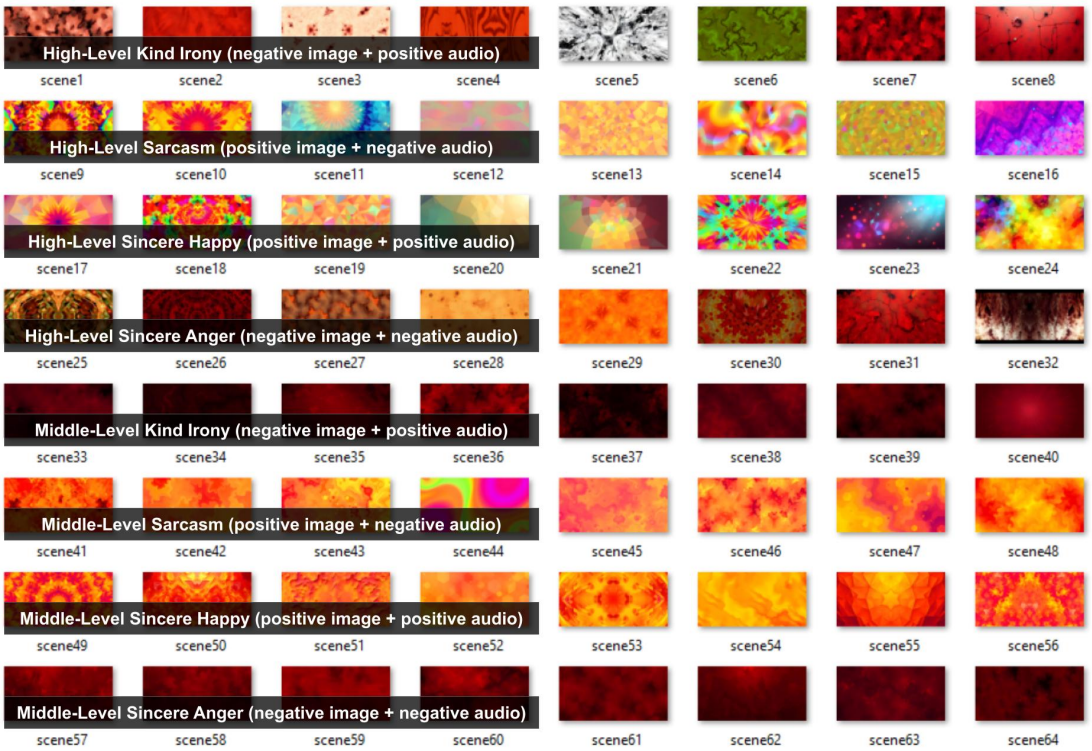


Figure 6.5: Visualization of the 64 generated environments used in the perceptual evaluation. The grid shows eight scenes for each audiovisual condition across two prompting strategies (high-level and mid-level). Conditions correspond to the proposed irony framework: Kind Irony (negative image + positive audio), Sarcasm (positive image + negative audio), Sincere Happy (positive image + positive audio), and Sincere Anger (negative image + negative audio).

component is synthesized using the MusicGen model. The resulting image and audio assets are automatically assembled into immersive audiovisual environments.

For the perceptual evaluation, eight representative audiovisual environments were selected, corresponding to the four affective configurations defined in this study: *Sincere Happy*, *Sincere Anger*, *Kind Irony* (negative image combined with positive audio), and *Sarcasm* (positive image combined with negative audio). Each configuration was generated using two prompting strategies: high-level prompts and mid-level prompts derived from statistically significant audiovisual descriptors identified in the previous analyses.

To present the stimuli to participants, a lightweight web-based visualization interface was developed using HTML and JavaScript. This interface enabled participants to interactively explore each 360° panoramic environment while listening to the corresponding soundtrack. The environments were hosted online and accessed through a standard web browser, allowing remote participation[2].

Participants evaluated the environments through a questionnaire implemented in Google Forms. After experiencing each scene, they rated the degree to which the audio matched the visual environment using a five-point Likert scale ranging from *Perfectly Matched* (1) to *Completely*

2. https://jfforero.github.io/AIS/

Mismatched (5). Higher scores therefore indicate stronger perceived audiovisual mismatch.

Because the responses were collected using an ordinal Likert scale and the sample size was limited, non-parametric statistical tests were used for all inferential analyses.

A total of 30 valid ratings were collected across the eight audiovisual environments. Participants evaluated the perceived alignment between the auditory and visual modalities using the five-point mismatch scale described above.

Table 6.4 summarizes the mean mismatch ratings for each scene.

Scene	Audiovisual Configuration	Prompt Level	Mean Rating
scene1	Kind Irony	High	3.75
scene2	Sarcasm	High	2.33
scene3	Sincere Happy	High	3.67
scene4	Sincere Anger	High	2.00
scene5	Kind Irony	Mid	1.33
scene6	Sarcasm	Mid	3.33
scene7	Sincere Happy	Mid	2.00
scene8	Sincere Anger	Mid	2.33

Table 6.4: Average perceived audiovisual mismatch scores for each generated environment. Lower values indicate stronger perceived alignment between modalities.

Results by Audiovisual Configuration To examine whether emotional opposition between modalities affects perception, ratings were aggregated according to the four audiovisual configurations defined in the proposed irony framework: *Kind Irony*, *Sarcasm*, *Sincere Happy*, and *Sincere Anger*.

Audiovisual Configuration	Mean Mismatch Score
Sarcasm	2.83
Kind Irony	2.71
Sincere Happy	2.63
Sincere Anger	2.22

Table 6.5: Average mismatch ratings grouped by audiovisual emotional configuration.

Overall, environments generated to express cross-modal emotional opposition (*Sarcasm* and *Kind Irony*) produced slightly higher mismatch scores than the sincere conditions.

To test whether this difference was statistically significant, ratings from ironic environments were compared with ratings from sincere environments using the Mann–Whitney U test. The analysis did not reveal a statistically significant difference ($U = 93$, $p = 0.39$). Although ironic environments tended to produce higher mismatch scores, the difference cannot be considered statistically significant given the available sample size.

Results by Prompting Strategy A second analysis examined whether the level of prompt specificity influenced perceived audiovisual alignment.

Environments generated using mid-level prompts produced lower mismatch ratings overall, indicating stronger perceived audiovisual coherence between modalities.

Prompting Strategy	Mean Mismatch Score
High-Level Prompts	3.00
Mid-Level Prompts	2.24

Table 6.6: Average perceived audiovisual mismatch by prompting strategy.

A Mann–Whitney U test comparing high-level and mid-level prompting strategies revealed no statistically significant difference ($U = 73$, $p = 0.07$), although the trend suggests that feature-derived prompts may improve audiovisual alignment.

Summary and Discussion

This chapter established a data-driven strategy for connecting measurable audiovisual features with affective prompt design for generative virtual environments. By synthesizing empirical analysis of musical and visual datasets with the historical and technical paradigms identified in the literature review, we developed a structured framework for cross-modal affective generation, in concordance with prior work in affective computing, audiovisual perception, and generative media systems.

Musical Feature Profiling

The statistical analysis of the *4Q Audio Emotion Dataset* demonstrated that emotional differentiation in music emerges from a combination of harmonic, dynamic, rhythmic, and timbral features, in concordance with established models of music emotion recognition and affective algorithmic composition (Livingstone et al. 2010; Kim and André 2004).

- **Arousal Indicators:** High-arousal emotions (*Angry*, *Happy*) were associated with increased onset density, higher loudness, and brighter timbral profiles (higher spectral centroid), aligning with prior arousal-energy mappings based on spectral brightness and acoustic intensity (Livingstone et al. 2010).
- **Valence Indicators:** Positive emotions (*Happy*, *Tender*) showed a strong prevalence of major tonalities, whereas *Angry* excerpts more frequently employed minor tonal modes, as seen in classical tonal-emotional correspondence theories; however, this relationship was not fully consistent across the dataset, particularly in rhythmically dominant excerpts where temporal density partially overrode harmonic modality.
- **Temporal Activity:** Onset rate proved to be a stronger indicator of emotional energy than global tempo, suggesting that rhythmic density contributes more directly to perceived intensity. This finding partially diverges from earlier rule-based systems that prioritize tempo as the primary temporal descriptor, indicating a shift toward micro-temporal structure as a more reliable predictor of perceived arousal.

Visual Feature Profiling

Analysis of the *Emotion6* dataset revealed that emotional differentiation in images is primarily encoded through chromatic and textural properties, in concordance with prior studies in visual affect and environmental psychology (Felnhofer et al. 2015; Marín-Morales et al. 2018).

- **Color and Luminance:** Positive emotions such as *Joy* are associated with higher brightness and saturation, whereas negative emotions such as *Fear* and *Sadness* exhibit darker and less saturated color distributions, aligning with established affective color theory; however, exceptions were observed in highly saturated low-valence scenes, suggesting contextual modulation of chromatic affect.
- **Visual Complexity:** High-arousal emotions are associated with more visually complex scenes (higher entropy and edge density), whereas calmer states tend to display smoother visual compositions. This is consistent with prior findings on perceptual load and emotional intensity in visual scenes, but not fully invariant across all categories, particularly in structured natural environments where complexity does not always correlate with arousal.

Prompt Engineering and Validation

The empirical insights were translated into a three-level prompt engineering framework: *High-level* (emotional labels), *Mid-level* (feature-derived descriptors), and *Low-level* (numerical parameters), in alignment with recent efforts to improve controllability in generative systems (Copet, Défossez, Gat, Nguyen, et al. 2023).

Perceptual validation showed that mid-level prompts produced lower mismatch scores than high-level labels, indicating stronger audiovisual coherence. This aligns with the hypothesis that feature-grounded descriptors provide better alignment between human interpretation and generative latent spaces. However, cross-modal emotional opposition (e.g., *Sarcasm*) produced higher mismatch ratings, as seen in prior studies on audiovisual incongruence and anempathetic sound (Chion 1994), but not uniformly across all participants, suggesting individual variability in interpreting emotional contradiction.

Chapter 7

Creative Practice

> *"Leí sus cuentos. Eran buenos. Leí sus poemas. Eran poemas mediocres, poesía mala como toda poesía escrita en un idioma que no es el nuestro."*
>
> *— Rim-Bot vuelve a casa*

Each iteration of the research process generated a series of technological artifacts that both supported and extended beyond the instrumental requirements of the affective virtual environment system. Some of these artifacts gradually evolved into autonomous creative works, functioning as independent projects with their own conceptual, aesthetic, and technical coherence. Collectively, they constitute a practice-based corpus organized around the conceptual axes of [no]-return and [no]-belonging.

The first two iterations of the project, Iterations 1 and 2, constituted an exploratory phase in which the core conceptual and technical frameworks of the research were established. These projects were developed in parallel with, and formally validated through, doctoral curricular units such as *Tópicos Avançados em Media Digitais* and *Seminário II*, undertaken during the 2020–2021 academic period under the residual and socially fragmented conditions of the COVID-19 pandemic.

Iterations 3, 4, and 5, grouped under the concept of *Emotional Machines*, incorporated an assemblage of computational mechanisms: real-time speech-to-text transcription, semantic-level sentiment analysis, acoustic speech emotion recognition, a musical progression generator based on the Tonal Interval Space (TIS), and a text-guided equirectangular image generation pipeline. Within this framework, we explored the transformation of immersive environments through spoken language. In most of the media projects developed during these iterations, participants were invited to speak in relation to a place or in response to the currently perceived 3D virtual space. These experiments resulted in a series of autonomous works developed within the *Emotional Machines* framework, including *Abandoned Memories*, *Veintitantos Vientos*, and *En train d'oublier*. In these works, speech functions as the primary interaction modality for spatial reconfiguration, where affective utterances reshape landscapes, atmospheres, and sonic structures in real time.

Iteration 6 extends the *AVE* generation system toward a specific geo-poetic and historical territory. In this iteration, the affective virtual environment is mobilized not merely as an abstract

emotional interface, but as a speculative cartographic device that revisits the historical imaginaries embedded in the notion of *Terra Australis Ignota*. This iteration was developed following the open call for works of the *XR Festival Florence* (2024), held at the *Cattedrale dell'Immagine* in Florence, Italy https://www.caphe.space/xr-festival-florence-2024. Curated as part of the festival's exploration of extended reality and spatial storytelling, the work was presented as an immersive internal mapping of the cathedral's architectural space (see Figure 7.1). A subsequent version of the project was exhibited in Lisbon during the *Noite Europeia dos Investigadores*, presented in an extended multi-surface projection format within the National Museum of Natural History and Science (see Figure 7.1). Presented in collaboration with academic institutions and museums, this iteration emphasized the project's hybrid positioning between research, artistic experimentation, and public engagement. Within this context, *Terra Australis Ignota* foregrounded the affective and speculative dimensions of scientific visualization, proposing emotional mapping as an alternative epistemic lens for engaging with territory, history, and memory.[1]

Iteration 7 deepens the investigation into emotional ambiguity arising from divergences between semantic sentiment analysis and acoustic speech emotion recognition. This ambiguity is exemplified in the *What Space* project, which emerged from observations of how prosodic variations in speech can produce emotional interpretations that differ from the literal semantic content of an utterance. This project was curated within the context of the *Loop Art Critique* residency and exhibition, hosted on the Onland.io metaverse platform (https://loop.onland.io/)[2].

Finally, through the deployment of the *AVE* system, it became possible to construct the conditions for generating environments whose audiovisual attributes could be juxtaposed and modulated to produce the *Ironic Machine[s]*: the *Sarcastic Machine* and the *Kind-Ironic Machine*. This technical development was first explored in Iteration 8 and further materialized in Iteration 9 through the artwork *Bello*, a multimedia project designed to speculate on the generation of affective virtual environments through speech emotion recognition in situations characterized by ambiguity and irony. The results of this iteration were later used to create a projection mapping installation at the Igreja da Sé in Faro, Portugal, presented in the context of the *Algarve Design Meeting* (see Figure 7.3).

Exhibitions and Presentations

- **Interactive Arts Installation at ACM Multimedia 2022 (Lisbon, Portugal):** Presented as part of the conference's multimedia exhibition, *Emotional Machines* demonstrated integrated VR rendering across multiple display contexts, including VR headset, dome projection, and 3D screen setups. The installation foregrounded embodied interaction through speech and joystick control, creating a space where emotional inference dynamically reshaped the environment.
  https://dl.acm.org/doi/10.1145/3503161.3549973

1. https://www.belasartes.ulisboa.pt/imersao-video-arte-na-investigacao-academica-noite-europeia-dos-investigadores/
2. Loop Art Critique is an experimental digital residency and critique program that supports a cohort of artists through structured critique sessions and culminates in a public exhibition of work within a shared virtual space

Figure 7.1: Public installations of *Terra Australis Ignota*. Top: immersive projection mapping inside the *Cattedrale dell'Immagine* during the *XR Festival Florence* (2024), where cartographic imagery and generated creatures were projected across the cathedral's architectural surfaces. Middle and Bottom: extended multi-surface projection installation presented at the *Noite Europeia dos Investigadores* at the National Museum of Natural History and Science, Lisbon, where the work was adapted to a gallery environment using panoramic projections across multiple walls.

Figure 7.2: Public exhibition flyer for the 13th Ambience on Onland.io. The What space.

- **Interactive Art Installation at Porto Tech Day 2023:** In this public technology and creativity event, the system was adapted to a screen-based setup with peripheral input devices (microphone, button, and keyboard). Users could search for geographic places, record audio descriptions, and observe how spoken language and emotions guided the generation of textured, affective visuals and associated music. This configuration emphasized accessibility and real-time engagement outside of VR headsets.
  https://portotech.scaleupporto.pt/

- **Demo Presentations:**

  – At CTM Inesc Day (Open Day CTM, INESC TEC), the project was showcased as a demonstration of affective virtual environment generation, highlighting the integration of machine learning techniques.

  – At the International Conference on Music AI (AIMC 2023), *Emotional Machines* was presented in the demo sessions under the theme of affective audiovisual systems, contributing to discussions around semantic conditioning of affective music generation and interactive AI instruments.
  https://aimc2023.pubpub.org/pub/chx6olpn/release/1

- **Digital and Visual Poetry Contexts:**

  – As digital poetry in the *Antología de Literatura Digital Latinoamericana*, where *Abandoned Memories* is framed alongside other works at the intersection of language, computation, and affective expression.
  https://antologia.litelat.net/volumen2/es/05_Abandoned_Memories.html

  – As visual poetry in the *Jornada de Poesía Visual*, where generated imagery and affective narratives were presented in dialogue with contemporary visual poetic practices.
  https://michaelmobius.github.io/muestra_jornadas/index.html

Figure 7.3: Architectural projection mapping of *Bello* at the Igreja da Sé, Faro, Portugal, presented during the *Algarve Design Meeting*.

Figure 7.4: Project installations at M.ou.co (top) and ACM (bottom), showcasing both screen-projection and immersive VR presentation modes.

7.1 Iteration 1: What Stirs Within

In the first iteration of the project, developed during the course *Advanced Topics in Digital Media* (2020-2021), we prototyped an interactive system that integrates speech, sentiment analysis, and real-time sound synthesis within a web-based environment. The implementation relied on JavaScript and the Web Speech API for speech-to-text transcription, enabling users to articulate short expressions in English that were then analyzed through a sentiment lexicon. Specifically, the system employed the AFINN emotional dictionary[3], a manually compiled list of 2,477 lexical entries annotated with integer values ranging from –5 (negative) to +5 (positive). Once a user's spoken phrase was recognized, the system calculated a cumulative emotional score and mapped it directly to the musical domain.

Sound production was handled through WebPD, a JavaScript library that ports Pure Data patches into the browser using the Web Audio API.[4]. This enabled sentiment values to dynamically control synthesis parameters in real time. Emotional polarity was musically articulated through tonal modality: positive sentiment activated major modes, whereas negative sentiment engaged minor modes. The auditory feedback consisted of a single sine-wave oscillator executing ascending or descending melodic contours depending on sentiment. To further expand the emotional palette, spectral experimentation was incorporated following (Wu, Falk, and Chan 2014), who associated instrumental timbres with affective categories (e.g., clarinet evoking sadness, saxophone and bassoon evoking joy). In this context, the system modulated synthesis parameters

3. F. Å. Nielsen, *A new ANEW: Evaluation of a word list for sentiment analysis in microblogs*. Available at: https://www2.imm.dtu.dk/pubdb/pubs/6010-full.html. Last accessed: March 9, 2026.

4. https://sebpiq.github.io/WebPd_website/ (Last accessed March 13, 2026).

such as modulation index, modulation frequency, and amplitude envelopes to simulate instrumental timbres. Randomized interval generation within modal frameworks, coupled with an expanded register of up to four octaves, contributed to a richer expressive range.

As illustrated in Figure 7.5 (left), the interface incorporated sliders that allowed users to manipulate synthesis parameters directly, adding a manual layer of control to complement the speech-driven interaction. Speech interaction was designed as an event-driven process rather than as continuous recognition. Although the Web Speech API allows persistent capture of user speech, the system intentionally operated in a single-utterance mode activated by explicit user input (click or tap). Each recognition cycle therefore unfolded as a discrete sequence: speech was triggered, transcribed, semantically classified, and immediately mapped to synthesis parameters. This turn-based design emphasized the performative intentionality of each vocal act.

Beyond sentiment mapping, the system also implemented rule-based voice commands. Keywords such as “hello world”, “music”, and “shut down” functioned as direct triggers, altering synthesis behaviors or interface states independently of emotional valence. In this way, the prototype operated simultaneously as an affective translator, mapping speech into emotional sound, and as a performative instrument, where speech could issue operational instructions. This dual interaction model foregrounded the ambiguity between expression and instruction, laying the groundwork for subsequent iterations of the project, in which the tensions between literal meaning and affective prosody would be further explored.

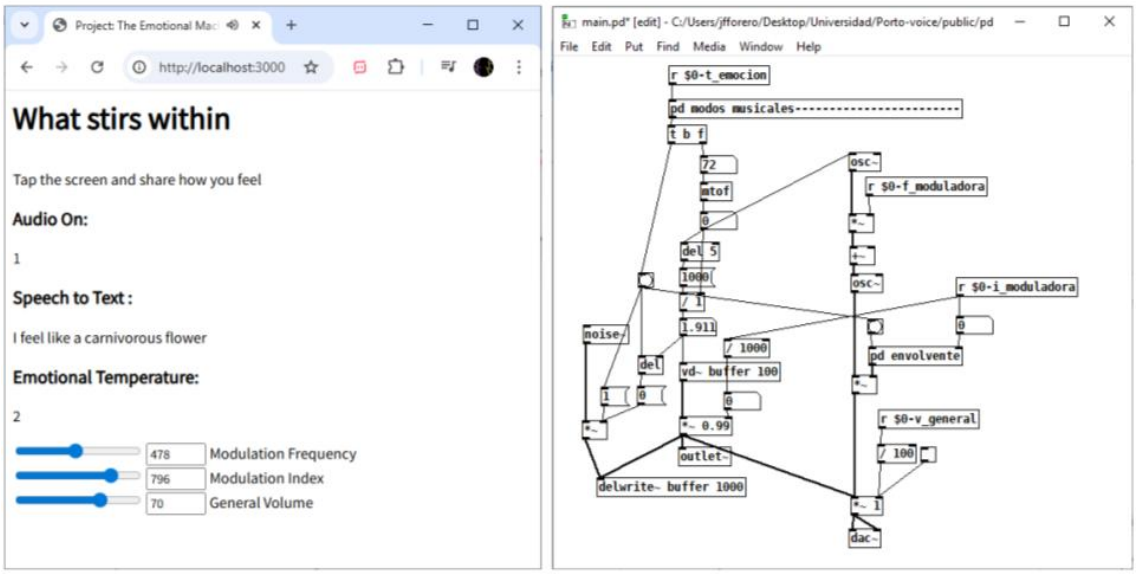


Figure 7.5: System architecture for real-time affective audio synthesis: (Left) Web-based front-end interface capturing speech transcription and user-defined emotional parameters; (Right) Corresponding Pure Data (Pd) patch performing dynamic audio modulation via the WebPd framework.

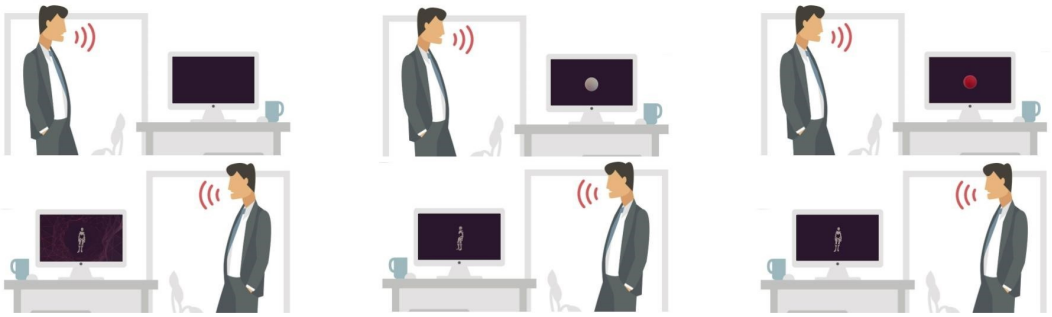


Figure 7.6: Voice-command interaction workflow in *A Máquina Sensível*. The user issues spoken commands that are transcribed through the speech recognition module and mapped to scene operations, enabling the real-time creation and transformation of virtual objects (e.g., spheres, humanoid figures, and color modifications).

7.2 Iteration 2: A Máquina Sensível

For the second iteration, we moved to *Unity3D*[5] for real-time 3D rendering power and efficient management of complex multimodal interactions. For speech-to-text, we used Unity's built-in Windows Speech API (`UnityEngine.Windows.Speech`) to achieve real-time, low-latency transcription. This audio was processed by a custom module (script) that manages dictation history and passes the finalized transcriptions to our analysis system.

We extended the command library to include a wider set of voice-driven interactions, covering creation, destruction, and modification of virtual elements in the scene. Virtual elements were divided into three main categories: Primitives, 3D Assets, and Computer Shaders. The set of primitives included a cube, a sphere, and a cylinder. The 3D assets used as examples were a humanoid character and a robot, while the generative shader effects implemented were contour lines and isolines, which could be dynamically enabled or disabled, as illustrated in Figure 7.7 (top).

Each element could be modified through standard transformation operations:

- **Translation**: up, down, left, right
- **Rotation**: clockwise, counter-clockwise
- **Scaling**: bigger, smaller
- **Colorization**: red, green, blue, random

Users can consult the available commands and interaction logic through a Help interface, shown in Figure 7.7 (bottom). In addition, the system allowed more advanced interactions, such as activating or muting music, changing the volume, controlling a virtual camera, triggering animations, and switching between narrative states.

5. Unity is a cross-platform real-time development environment widely used for interactive 3D graphics, simulation, and game development. https://unity.com/ (Last accessed March 13, 2026).

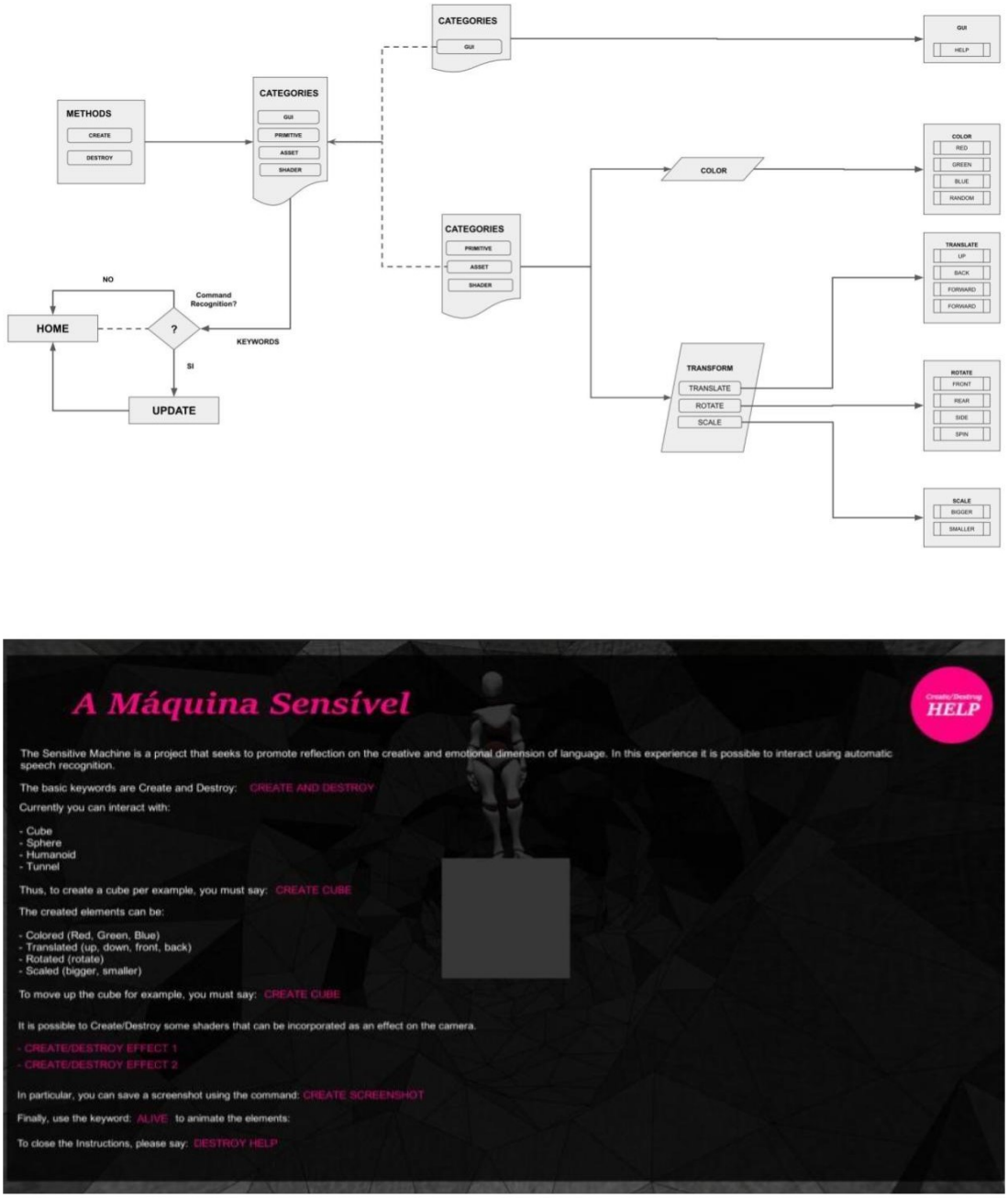


Figure 7.7: System architecture of *A Máquina Sensível*. (Top) Command-recognition logic showing the hierarchical organization of methods, categories, and keywords used to interpret spoken instructions. (Center) Transformation and attribute operations available for virtual objects, including colorization, translation, rotation, and scaling. (Bottom) Graphical Help interface presenting the available voice commands and interaction instructions for the user.

Language

Language is _____, no longer what it used to be,
the heroic dream of the future.

The suffering organisms die.
And if they do not die, then—

Why should their bodies exist?

The emotional machines in the sky are as brutal as they ever were.
When they began to churn, the sun started to pout.

Language is _____, no longer what it used to be,
the heroic dream of the future.

7.3 The Emotional Machine[s]

We integrated the DeepAI API[6] to generate and manipulate images from text prompts, further enhancing them through services such as image upscaling and neural style transfer. As shown in Figure 7.8, the processed images were mapped as textures onto a 3D environment constructed from basic geometric primitives (planes, cubes, and spheres). To facilitate interaction, we developed a custom graphical user interface that enabled speech-based activation and real-time control of the image generation process. For immersive deployment, the system was executed on an Oculus Quest 2 headset with joystick input. The joysticks supported a first-person navigation system, allowing users to translate, rotate, and interact with the virtual space via a laser-pointer selection tool. User interaction was time-constrained, granting up to six seconds per activation to issue spoken input through the headset's integrated microphone. The transcribed speech was also represented in the environment as a 3D Text object.

Figure 7.8: Screenshot of the Affective Virtual Environment (AVE) system developed between 2021 and 2022. The image shows the result of a user wearing a VR headset speaking the phrase "we were young and we were crazy". The generated image derived from this text prompt is subsequently used as a style source to transform a second image (an equirectangular panorama) selected according to the sentiment analysis of the spoken input. The interface also displays the GUI with two VR buttons ("SPEAK" and "GENERATE") and the real-time transcription of the user's speech.

At the semantic layer, as presented in this research, we introduced a Rocchio classifier for ternary sentiment analysis (positive, negative, neutral). For this instance, the model was trained

6. DeepAI provides machine learning APIs for tasks such as image generation, enhancement, and style transfer using deep neural networks. See https://deepai.org/ (last accessed March 13, 2026).

on a manually curated dataset[7] IMDB of English-language movie reviews and social media posts. Texts were preprocessed and vectorized using a TF-IDF weighting scheme, yielding prototypical concept vectors for each sentiment class. These vectors were distilled into interpretable lexicons of weighted terms, enabling the runtime system to perform rapid lexicon-based sentiment scoring without the computational overhead of a full vector-space model.

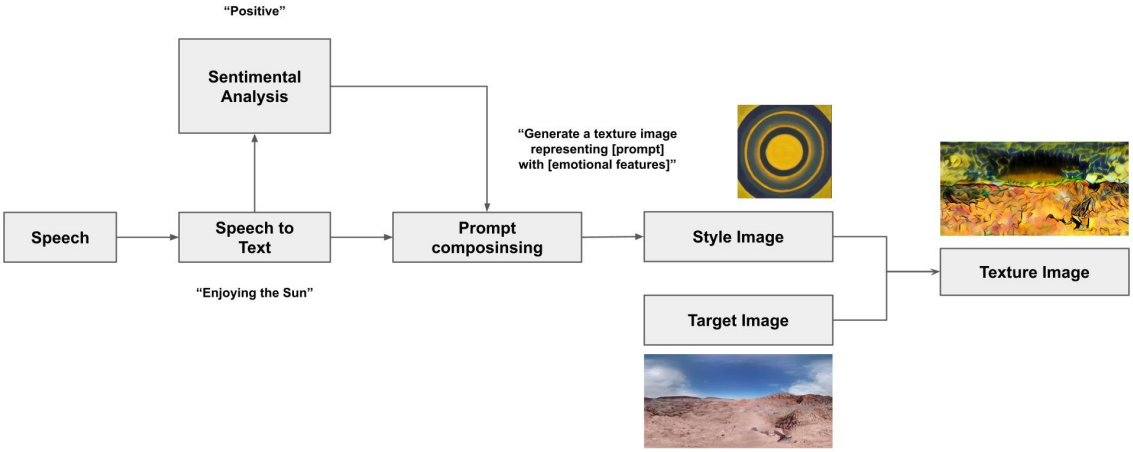


Figure 7.9: Flowchart for affective texture synthesis: The workflow integrates speech-to-text transcription with sentiment analysis to compose generative prompts, which are then used to produce style images for final equirectangular target image transformation.

We were particularly interested in images for sphere texturing. Since the images generated were not natively adapted for equirectangular projection, we introduced a pipeline (see Figure 7.9) in which the predicted sentiment values guided the selection of a curated equirectangular image. This selected image served as the structural target, while the image generated from the original text prompt functioned as the style source. A style transfer process was then applied, merging the semantic alignment of the curated projection with the aesthetic qualities of the generated output (see Figure 7.10). This approach ensured that the final immersive textures were both geometrically coherent for spherical mapping and emotionally consistent with the sentiment layer, enabling seamless integration into the virtual 3D environment.

Generation of Equirectangular Panoramas from Google Street View

We developed a computer program in C# to automatically generate equirectangular panoramas from specific geographical locations within the Unity environment.[8] This process leverages the Google Street View Static API to extract six distinct images corresponding to the six faces of a

7. The IMDB Large Movie Review Dataset is originally annotated with binary sentiment labels (positive and negative). In our implementation, the dataset was used to estimate word-level polarity weights by computing the relative frequency of terms in positive and negative reviews. Although the training data is binary, the system can produce a neutral outcome when the aggregated lexical evidence does not strongly favor either polarity, allowing the classifier to operate effectively in a ternary sentiment framework.

8. Implementation available in the public repository https://github.com/jfforero/Map2Unity/ (Last accessed March 13, 2026).

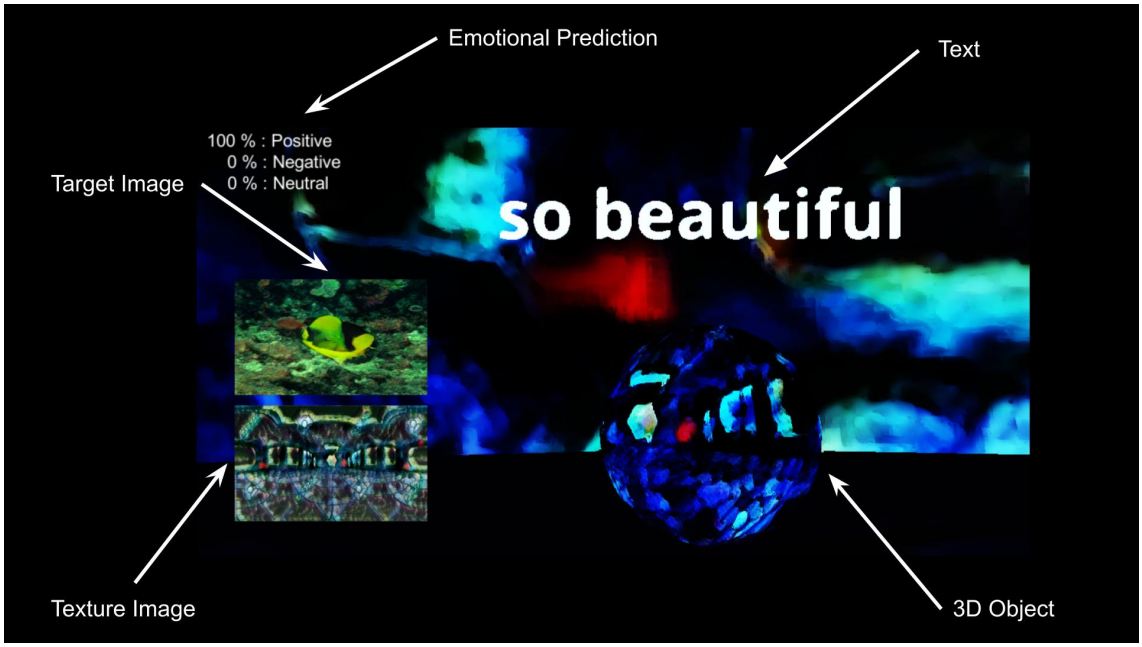


Figure 7.10: Interface overview of a synthesized Affective Virtual Environment (AVE). The screenshot displays the real-time mapping of speech-derived sentiment onto a 3D object's texture, illustrating the convergence of text, target imagery, and generated emotional profiles.

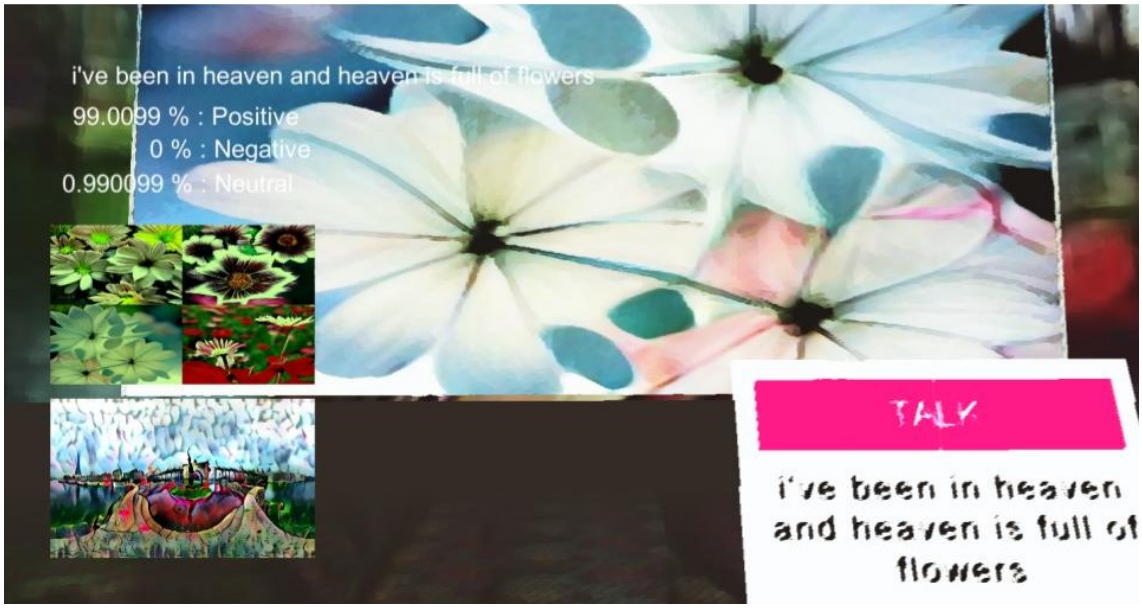


Figure 7.11: Updated Affective Virtual Environment user interface. The system demonstrates the real-time integration of speech-to-text transcription, sentiment analysis, and the resulting generative texture synthesis applied to the immersive 3D scene.

cube: front, back, left, right, top, and bottom. These views are obtained by varying the orientation parameters of the API call, specifically the *heading* (for horizontal angles) and the *pitch* (for vertical angles). The four horizontal views are generated by rotating the heading angle at 90-degree intervals, while the top and bottom views are captured by adjusting the pitch to $+90^\circ$ and -90°, respectively.

Once the six directional views are retrieved, they are arranged into a composite cubemap layout in a 4×3 grid structure. This layout serves as the input for a custom transformation algorithm that converts the cubemap into an equirectangular projection. [9] The conversion routine is implemented as a static class in Unity and operates by iterating over the spherical coordinates of the target equirectangular image. Each pixel is mapped from angular coordinates (ϕ, θ) into a corresponding 3D unit vector. This vector is then projected onto the appropriate face of the cubemap, where bilinear interpolation is used to sample the corresponding color. The final output is a panoramic image with a $360^\circ \times 180^\circ$ field of view, formatted in the equirectangular projection (see Figure 7.12).

The Extended Conchord

Predicted emotional states mapped to musical properties through an extended version of *Conchord++*, a real-time harmonic generation system implemented in *Pure Data* (Pd) and grounded in the Tonal Interval Space (TIV) framework.[10] In this system, affective predictions obtained from the multimodal analysis directly modulate harmonic, spatial, and temporal parameters, allowing emotional inference to be expressed as structured musical behavior.

Positive emotional predictions are mapped to regions of high consonance within the TIV, corresponding to vectors with large magnitudes and low spatial dispersion. Harmonically, this results in stable tonal configurations and smooth trajectories that follow circular motion along the circle of fifths, reinforcing perceptual coherence and tonal proximity. These trajectories generate on-tempo rhythmic sequences with higher event density, producing a sense of regularity and forward momentum. For positive states, tempo oscillates within a predefined higher range (between values c and d), reinforcing energetic and affectively congruent musical motion.

Conversely, negative emotional predictions are mapped to regions of lower consonance and higher dispersion in the TIV, yielding vectors closer to the tonal center and traversing more distant harmonic relationships. In these cases, harmonic motion adopts rectilinear trajectories that cut across the circle of fifths rather than circulating within it, producing abrupt or tension-oriented progressions. Rhythmic structures become sparser, with reduced note density and irregular temporal articulation. Tempo oscillates within a lower and less stable range (between values a and b), contributing to a perceptual sense of instability and emotional friction.

9. Adapted from open-source implementations for cubemap-to-equirectangular projection, https://github.com/Mapiarz/CubemapToEquirectangular (Last accessed March 13, 2026).

10. Tonal Interval Space (TIV) is a multidimensional representation of tonal relationships derived from the Fourier transform of pitch-class sets, widely used in computational music analysis and algorithmic composition. See https://sites.google.com/site/tonalintervalspace/home (Last accessed March 13, 2026).

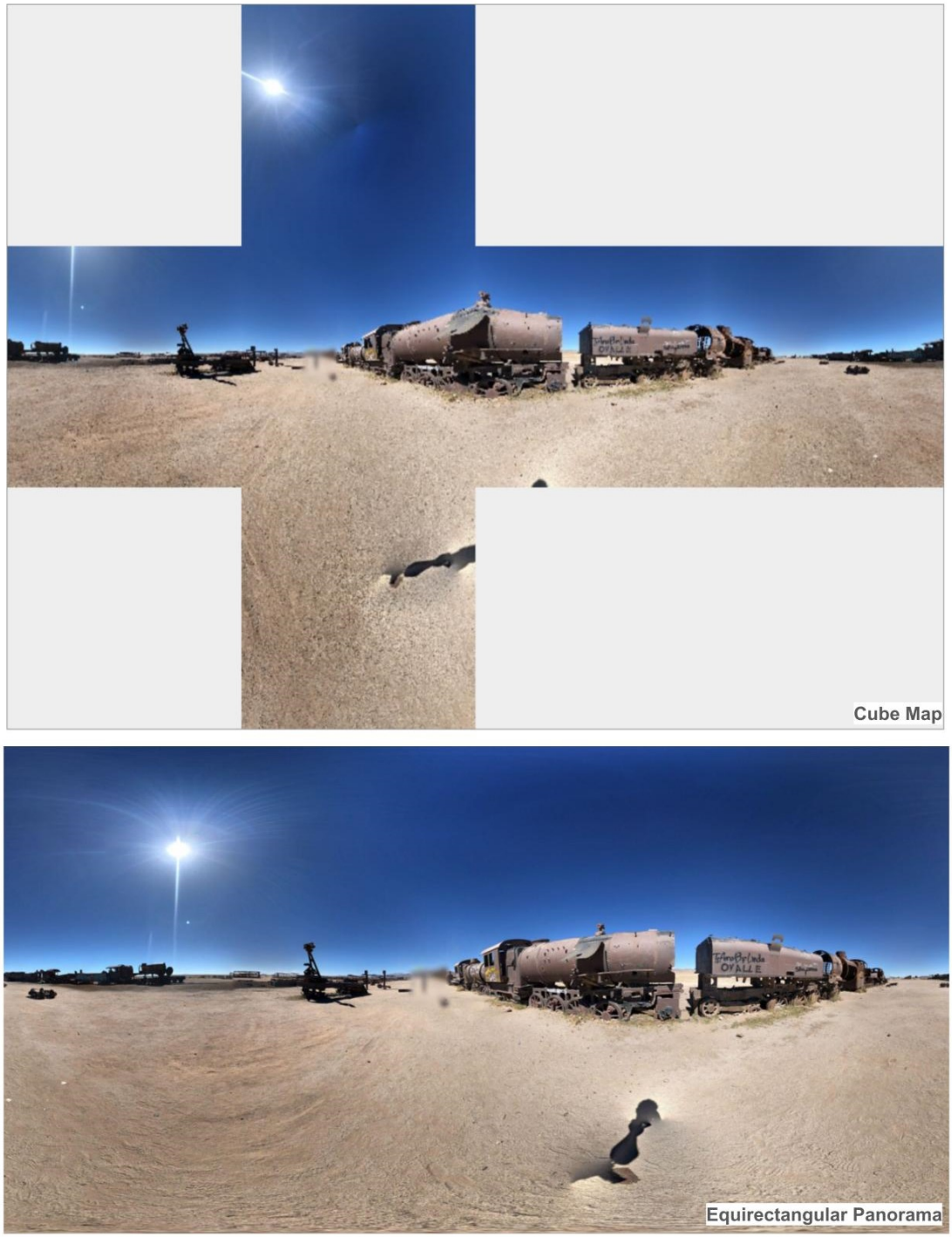


Figure 7.12: Mapping of 3D spatial data for immersive environments: (Top) Six-face Cube Map representation of a virtual scene; (Bottom) The resulting Equirectangular Panorama used as the target image for affective style transfer.

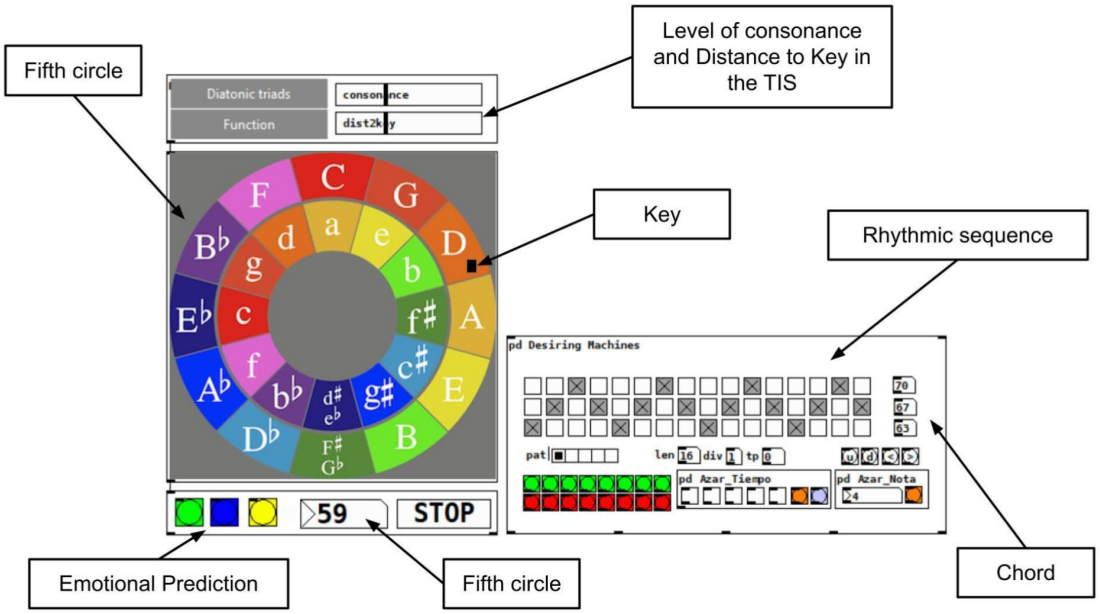


Figure 7.13: Conchord Extended: Graphical user interface description for Music generation using the tonal interval space in Pure data

For emotional categories outside the positive–negative polarity, such as neutral or ambiguous affective states, the system adopts an intermediate strategy. The tonal center remains fixed, avoiding large-scale harmonic modulation, while tempo oscillates within a mid-range interval (between values e and f). Rhythmic sequences are intentionally off-tempo, introducing temporal misalignment without overt harmonic tension. This configuration reflects affective ambiguity by maintaining tonal stability while destabilizing rhythmic expectation.

Through this mapping strategy, the extended Conchord system translates affective predictions into navigable trajectories within the Tonal Interval Space, where consonance, harmonic distance, rhythmic density, and temporal variation jointly encode emotional qualities. The result is a perceptually grounded, musically interpretable soundtrack that bridges affective computing and generative composition. The system ouput was conected via MIDI to AbletonStudio to explore diferent synthetic textures used for different presentation and demos.

7.3.1 Iteration 3: Abandoned Memories

We were particularly interested in the possibility of visiting unknown places (at least for the author, in the sense of not having been physically visited) and experiencing a sense of emotional resonance within them. This exploration led us to reflect on nostalgia or longing for a time or place never personally lived, a feeling that John Koenig described as *anemoia*[11]. The idea was permit participants to speak about the emotions evoked by the virtual environment, and to use these spoken words to generate an image that would serve as the style source for transfer onto

11. The term *anemoia* was coined by John Koenig to describe "nostalgia for a time you have never known." See John Koenig, *The Dictionary of Obscure Sorrows* (New York: Simon & Schuster, 2019).

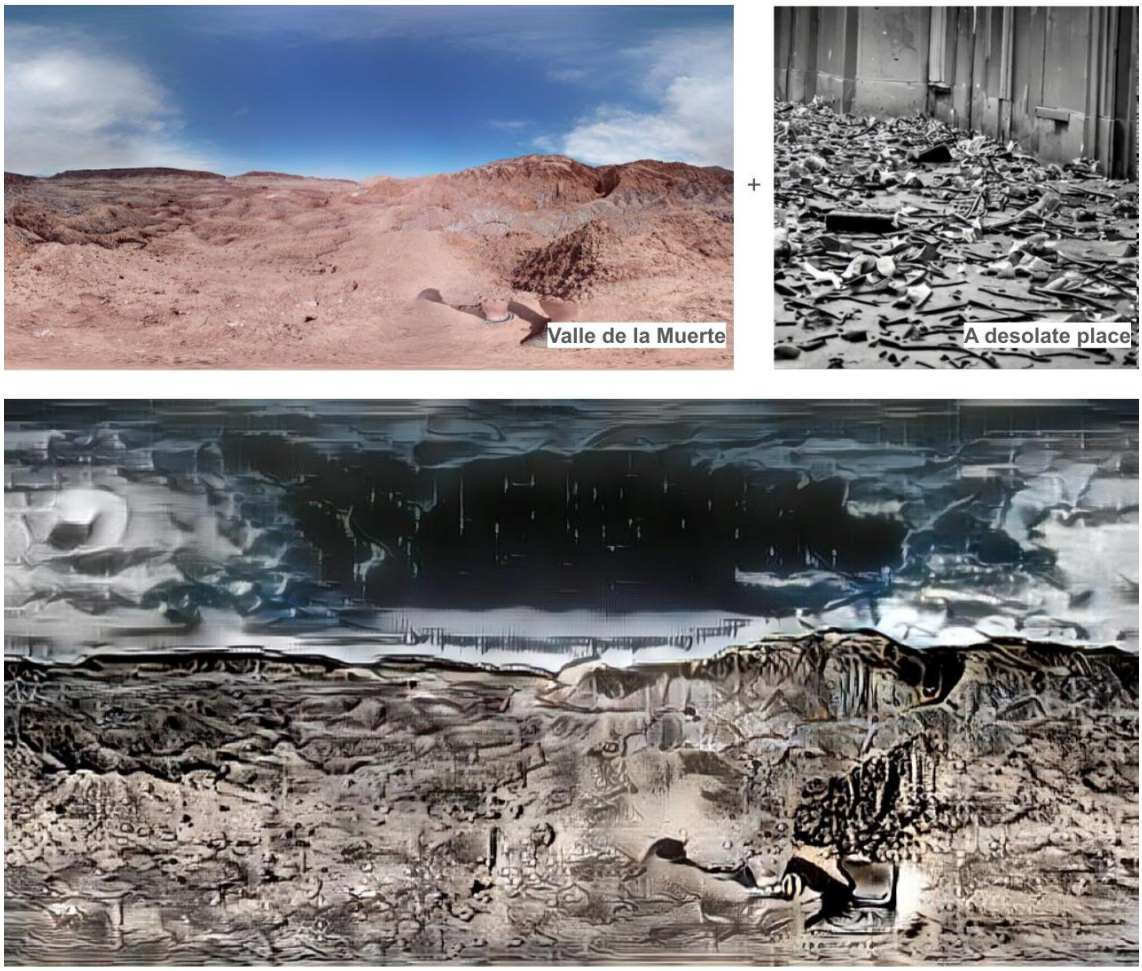


Figure 7.14: Affective texture generation via Neural Style Transfer. The process demonstrates the transformation of a base equirectangular panorama of "Valle de la Muerte" (left) into a stylized texture guided by the emotional prompt "desolate place" (right), resulting in a visually congruent environment for low-valence affective states.

a predefined equirectangular image, thereby transforming the virtual space. We selected a series of locations to construct an imaginary of remote and abandoned environments. In particular, we were drawn to places such as Valle de la Muerte and the Pukará de Quitor, both located in the Atacama Desert in northern Chile, whose landscapes evoke strong associations with disappearance, abandonment, and loss.

7.3.2 Iteration 4: Veintitantos Vientos

We further deployed the positioning system across 24 locations in the surroundings of the author's residence in Porto, Portugal. At each site, reflections related to the environment were articulated, including the recitation of fragments from the poem *Veintitantos Vientos*.

VEINTITANTOS VIENTOS

Bajamos a la orilla del río,
al río,
donde está la cascada—
y no vimos agua
ni nubes.

Me pareció sentir
un mar vacío.
Me pareció navegar
por un mundo muerto.

Entonces te veo,
sonriéndome y riéndote,
y te ríes—
y luego me miras,

y te miro—
y pienso,
por un segundo,
que no es necesario volver a casa.

Se nos ha hecho de noche;
llueve, y hace frío.
Tengo la sensación
de que solo vivimos
en las ocasiones más pequeñas.

Debemos subir
y no quiero.
El jardín de San Lázaro
no está lejos.
Caminamos cogidos del brazo—
un gesto que nos tomamos muy en serio.

Luchamos contra la muerte,
de forma sutil,
aunque de manera inútil
y hasta a veces, peligrosa.

Las nubes nos detienen.

Teníamos veintitantos vientos
cuando nos conocimos,
y en Campo 24 de Agosto
dejamos nuestras vidas.

Teníamos veintitantos vientos.
Veintitantos vientos.
Tanto, tanto.[12]

The objective was to produce multiple variations of a single motif, shaped by the characteristics of each location and the wording of the recited poem. A user interface was developed to allow the input of geographic positions through text commands, which were then retrieved via Google Street View and represented within our immersive system. The speech-to-text transcription was post-processed to generate more structured prompts, and the neural style transfer algorithm was replaced with a diffusion-based model capable of performing semantic image editing.

In this way, a transcription such as "I felt an empty sea, I sailed through a dead world" was automatically reformulated into a more structurally precise prompt, for example: "transform this image to represent the idea of [transcription]". In this context, "this image" refers to the selected location—such as the Coreto de São Lázaro in Porto, Portugal—as shown in Figure 7.15.

The objective was to produce multiple variations of a single motif, shaped by the characteristics of each location and the wording of the recited poem. A user interface was developed to allow the input of geographic positions through text commands, which were then retrieved via Google Street View and represented within our immersive system. The speech-to-text transcription was post-processed to generate more structured prompts, and the neural style transfer algorithm was replaced with a diffusion-based model capable of performing semantic image editing.

In this way, a transcription such as "I felt an empty sea, I sailed through a dead world" was automatically reformulated into a more structurally precise prompt, for example: "transform this image to represent the idea of [transcription]". In this context, "this image" refers to the selected location—such as the Coreto de São Lázaro in Porto, Portugal—as shown in Figure 7.15.

Twenty-four locations surrounding Praça de São Lázaro and the Estação 24 de Agosto were represented as affective environments. The resulting graphical outputs extend the equirectangular

12. The poem *Veintitantos Vientos* forms part of the online poetic project *Rim-Bot vuelve a casa*. See https://rimbotvuelveacasa.wordpress.com/.

panorama and the VR headset experience, composing a visual poem (see Figure 7.16) that encodes fragments of the poem within each generated image.

7.3.3 Iteration 5: En train d'oublier

Iteration 5 deepens the project's engagement with the poetics of *anemoia* and the possibility of experiencing the unknown. This iteration was situated among the abandoned trains on the outskirts of the Salar de Uyuni in Bolivia, a site commonly known as the *train cemetery*. These locomotives were originally deployed in the early twentieth century to support mineral transport across the Bolivia–Chile corridor, but were progressively abandoned following shifts in extraction economies and the regional decline of railway infrastructure during the 1930s. Over time, the vehicles became both debris and monument: an industrial remainder that stages memory as a material condition rather than a narrative certainty. Once situated, we uttered short sentences imagining possible (or impossible) futures, micro-fictions, descriptions, and affective guesses. These utterances were captured, transcribed, and used to generate multiple affective virtual environments, as before, from which we extracted audiovisual material to compose a set of techno-poetic artifacts.

Using video and web-art techniques, we experimented with the generated panoramas to produce binary and cyclical visualization proposals.[13] The binary procedures emphasize discontinuity through a web-based interface where two instances of the same generated panorama can be compared interactively using a sliding divider (Figure 7.17). In contrast, the cyclical procedures foreground duration and return: a continuous video loop in which one generated image slowly transforms into another and eventually returns to its initial state, producing a visual rhythm that never fully coincides with itself.

The work stages belonging as a grammatical performance through the conjugation of the verb *être* (to be). The poetic gesture consists in completing each conjugated form with the work's title—*en train d'oublier*. In doing so, identity becomes an unfinished predicate: being is articulated as a temporal drift rather than a stable state—je suis en train d'oublier, tu es en train d'oublier, il est en train d'oublier... In French, the word *train* literally means "train," but in the expression *être en train de*, it designates an ongoing process or action.

13. Interactive web version available at https://jfforero.github.io/En-train-d-oublier-web-edition-/. A cyclical video-loop version of the work can be viewed at https://youtu.be/tz9Lhuyw_sw.

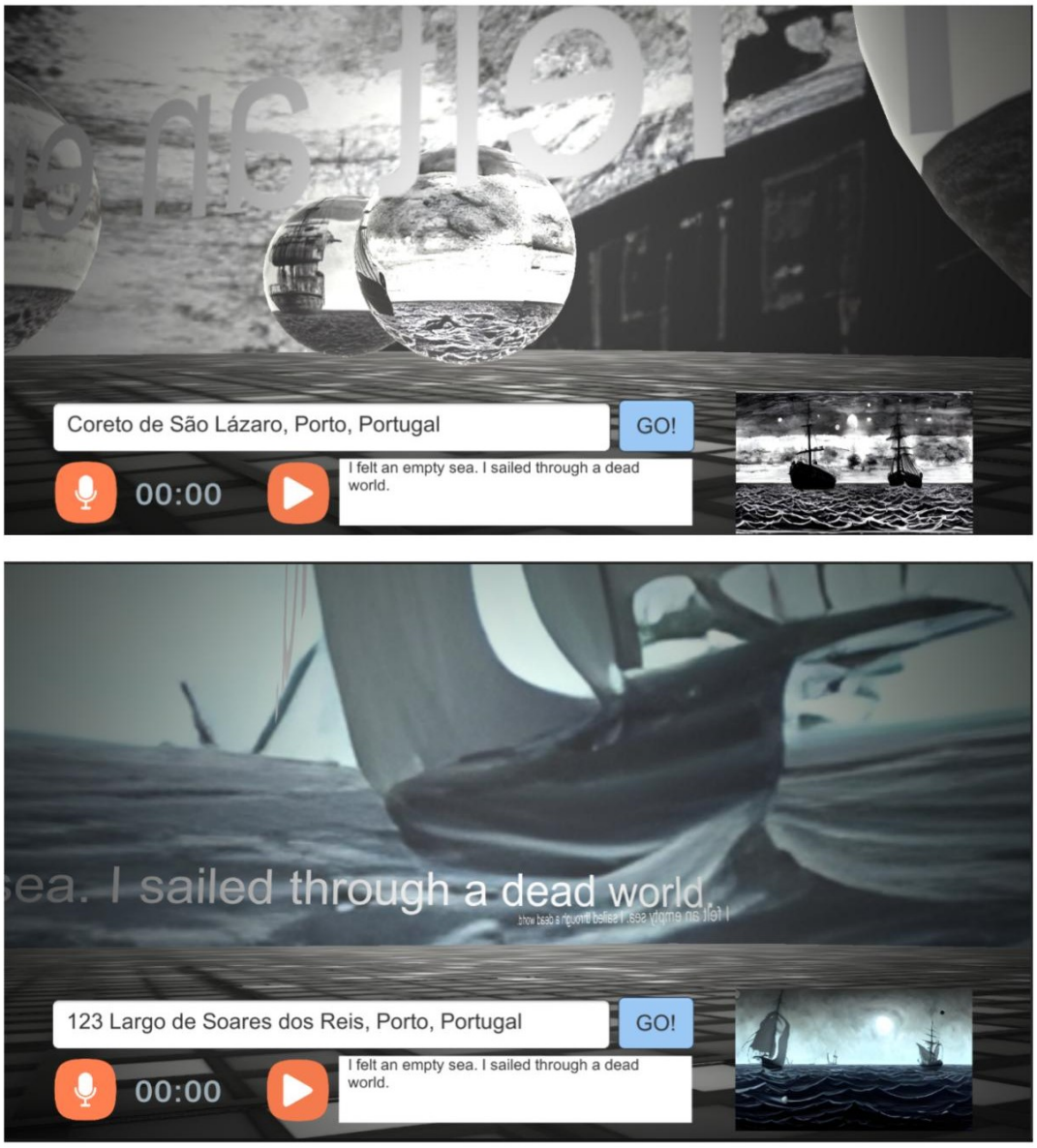


Figure 7.15: The *Emotional Machines* interactive interface. The GUI facilitates the input of geographic coordinates alongside real-time voice recording to synthesize immersive equirectangular panoramas through affective image diffusion and text-guided stylization.

Figure 7.16: The visual poem *Veinte y tantos Vientos*: A collection of 24 equirectangular panoramas generated using the *Emotional Machines* system. Each image represents a distinct geospatial coordinate within the Campo 24 de Agosto area in Porto, Portugal, synthesized through text-guided affective diffusion.

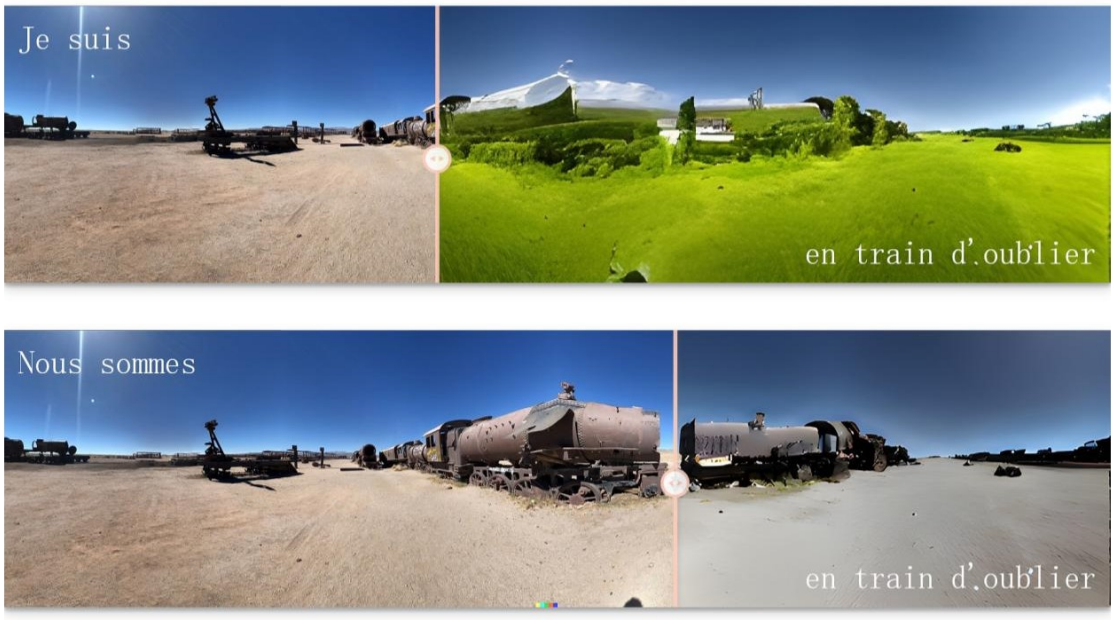


Figure 7.17: Binary interface from the web-based artwork *En train d'oublier*. Each panorama is split by an interactive vertical slider that allows viewers to reveal two simultaneous states of a generated environment. On the left, the landscape corresponds to the abandoned railway site at the Salar de Uyuni, while on the right the same environment is algorithmically transformed into speculative or impossible futures. The conjugated forms of the verb *être* ("Je suis," "Nous sommes") are paired with the phrase *en train d'oublier*, framing identity as an ongoing process of becoming and forgetting.

EN TRAIN D'OUBLIER

je suis

tu es

il est

nous sommes

vous êtes

elles sont

en train d'oublier

en train d'oublier

en train d'oublier

Figure 7.18: The visual poem *En train d'oublier*: a collection of equirectangular panoramas generated using the *Emotional Machines* system. Each image is a distinct transformation of the same site through text-guided affective diffusion.

7.4 Iteration 6: Terra Australis Ignota

As part of our technical exploration in this iteration, we integrated our progress in the acoustic-based Speech Emotion Recognition system exposed in 4.2. For speech-to-text transcription, we adopted *Whisper*[14], which provides multilingual semantic capabilities even though the underlying training of our acoustic model was conducted exclusively in English. For the SER component, we implemented a lightweight LSTM architecture trained on 40 Mel-Frequency Cepstral Coefficients, extracted with *Librosa* from the RAVDESS emotional speech database. The network was optimized using the RMSprop optimizer with categorical cross-entropy loss.[15].

The term *Terra Australis Ignota* refers to a set of cartographic hypotheses developed between the 15th and 18th centuries, through which European navigators and cosmographers postulated the existence of vast southern lands largely unknown to them. Regions such as Patagonia, Tierra del Fuego, and Antarctica were represented not through empirical certainty, but through conjecture, narration, and extrapolation. These territories were frequently depicted as exotic, hostile, or monstrous, populated by giants, hybrid beings, and fantastical fauna—graphic projections that revealed as much about European epistemologies as about the lands themselves.

Following the arrival of Columbus in the Americas and the subsequent establishment of the Casa de Contratación in 1503, cartographic production intensified dramatically. The *Padrón Real*, a continuously updated and confidential royal map, centralized geographic knowledge gathered from explorers' accounts. Yet these maps remained inherently incomplete. Large portions of the southern hemisphere were rendered as blank, fragmented, or ambiguously annotated spaces, often accompanied by illustrations of flora, fauna, and mythological creatures. The unknown thus became visually and narratively coded as monstrous, unstable, and excessive.

Iteration 6 appropriates this historical condition of cartographic indeterminacy as a conceptual and technical framework. By situating the Emotional Machines system within southern Chilean landscapes, the project reactivates the tension between documentation and imagination, between geographic specificity and affective projection. Speech—captured, transcribed, emotionally analyzed, and transformed through generative systems—functions here as a contemporary analogue to the explorers' testimonies: partial, subjective, and affect-laden accounts that reshape the representation of territory.

We uploaded this version to Hugging Face and deployed it in two distinct interactive interfaces. Through these two interfaces, developed with Gradio[16], we explore how acoustic emotion recognition and speech transcription can be merged with generative imagery to create affective virtual environments, enabling a multimodal interplay between sound, language, and visual imagery.

14. Speech transcription was implemented using the `faster-whisper` library, an optimized implementation of OpenAI's Whisper automatic speech recognition model. https://github.com/openai/whisper and https://github.com/guillaumekln/faster-whisper.

15. The training implementation is available as a Google Colab notebook: https://github.com/jfforero/The-Ironic-Machines/blob/main/SER-LSTM-RAVDESS-40MFCC-TRAIN.ipynb

16. The system was deployed using Hugging Face Spaces, a platform for hosting machine learning applications, and Gradio, an open-source Python library for building interactive web interfaces for ML models. See https://huggingface.co/spaces and https://www.gradio.app.

1. **Emotional Maps Generation**

   In this instance, speech is used to construct a multimodal prompt by combining automatic speech-to-text transcription with acoustic emotion prediction derived from prosodic features. The resulting prompt integrates both semantic content and affective state. This prompt is applied to a curated collection of equirectangular images depicting extreme South American landscapes. At each iteration, a base image is randomly selected from this dataset and submitted, together with the generated prompt, to an external image transformation model via API. The transformation is guided by the following instruction:

   > Transform this image into an ancient exploratory map of an unknown land. The territory emerges from: *[transcribed_text]*. Its geography, symbols, and distortions should evoke the emotion of *[emotion_prediction]*.

   The generative process reinterprets the original landscape as a speculative cartographic surface, where geography becomes modulated by both linguistic description and vocal affect. This results in a series of "emotional maps," extending the notion of *Terra Australis Ignota* as an imagined and affectively constructed territory.[17]

2. **Emotional Monsters Generation**

   In contrast, this second instance bypasses curated images and directly generates visuals from the textual prompt itself. The output is a synthetic representation of "Patagonian Monsters" whose appearance is modulated by the predicted emotion and transcribed utterance.[18]

   > Generate Patagonian Monsters with a *[emotion]* attitude, representing the idea of: *[transcribed text]*. Illustrate this using asemic writings in an old map style.

The material produced within this frameworks was subsequently used to develop an immersive installations presented in two distinct exhibition contexts. The first took place at the *XR Festival Florence* (2024), held at the *Cattedrale dell'Immagine* in Florence, Italy, where the system was deployed as an internal projection mapping of the cathedral's architectural space. We complemented the material with shaders graphic animations made to simulate la cordillere and the sea. A subsequent version was adapted for presentation in Lisbon during the *Noite Europeia dos Investigadores*, where the work was reconfigured into an extended multi-surface projection environment within the National Museum of Natural History and Science. The generated visual material was complemented by shader-based graphical animations designed to evoke the Andes mountain range and the sea.

17. The Emotional Maps interface is available at: https://huggingface.co/spaces/jfforero/Terra_Autralis_Ignota. Accessed: March 2026.
18. The Emotional Monsters interface is available at: https://huggingface.co/spaces/jfforero/Terra-Autralis-Ignota. Accessed: March 2026.

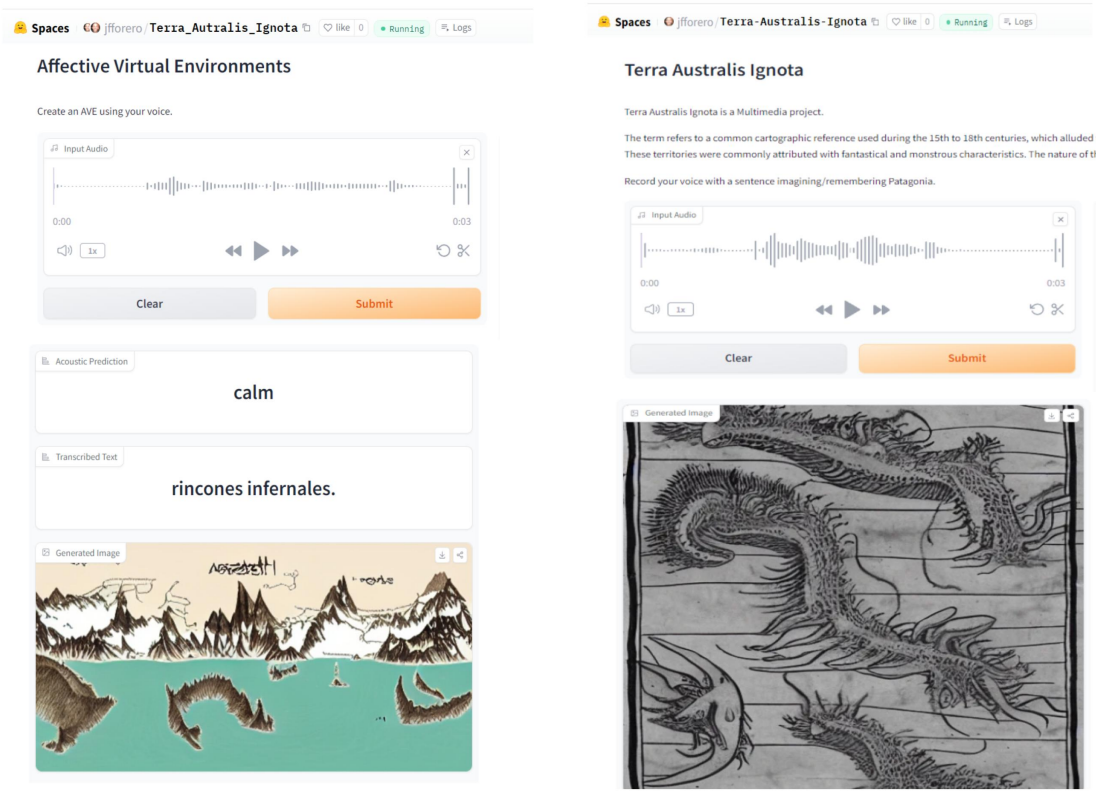


Figure 7.19: Left: Example of Emotional Maps generation based on the fusion of emotional speech and curated cartographic imagery. Right: Example of Emotional Monsters generation, where emotion and language drive the direct creation of synthetic creatures.

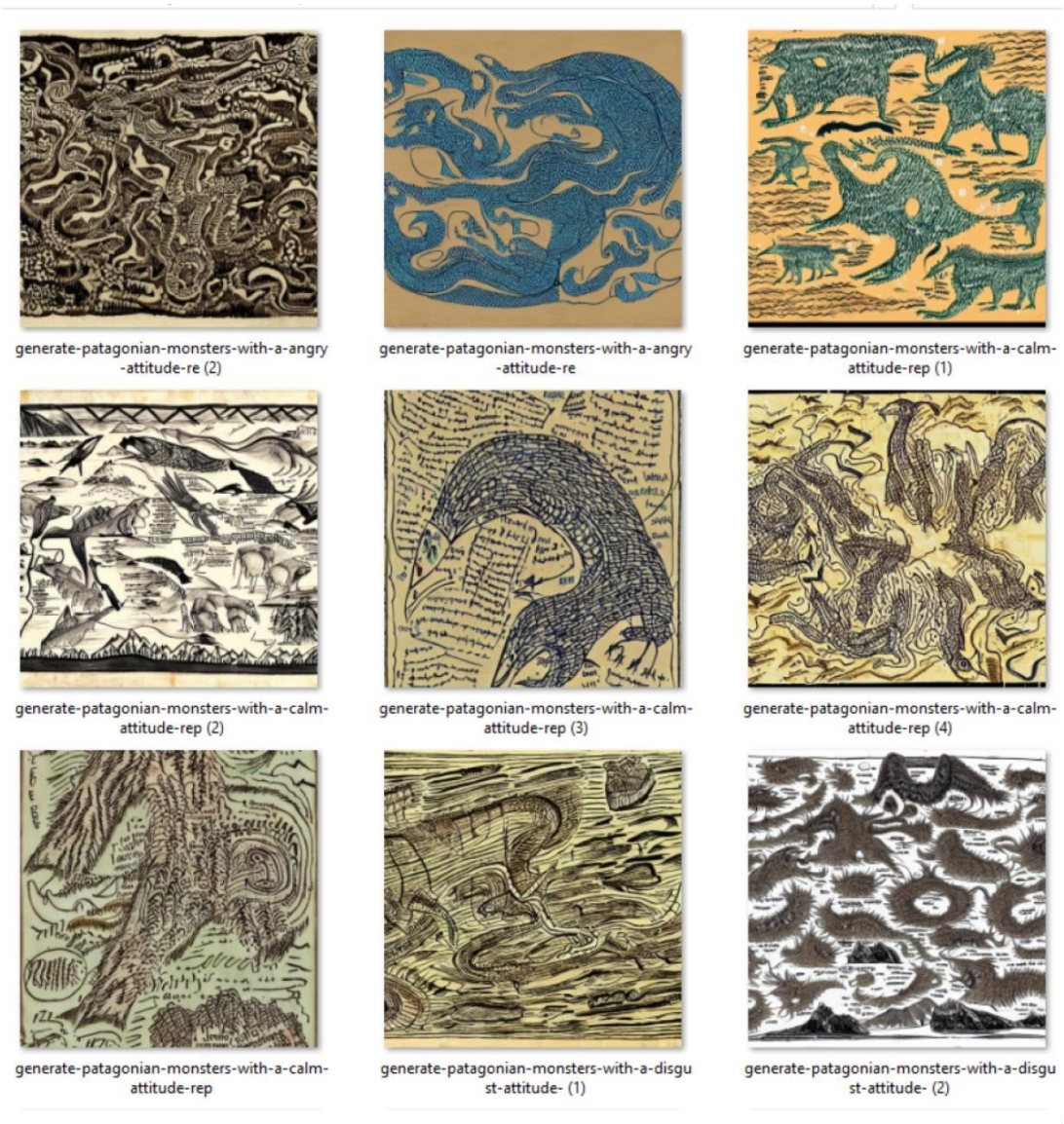


Figure 7.20: Folder containing image files created with the Monster generator system.

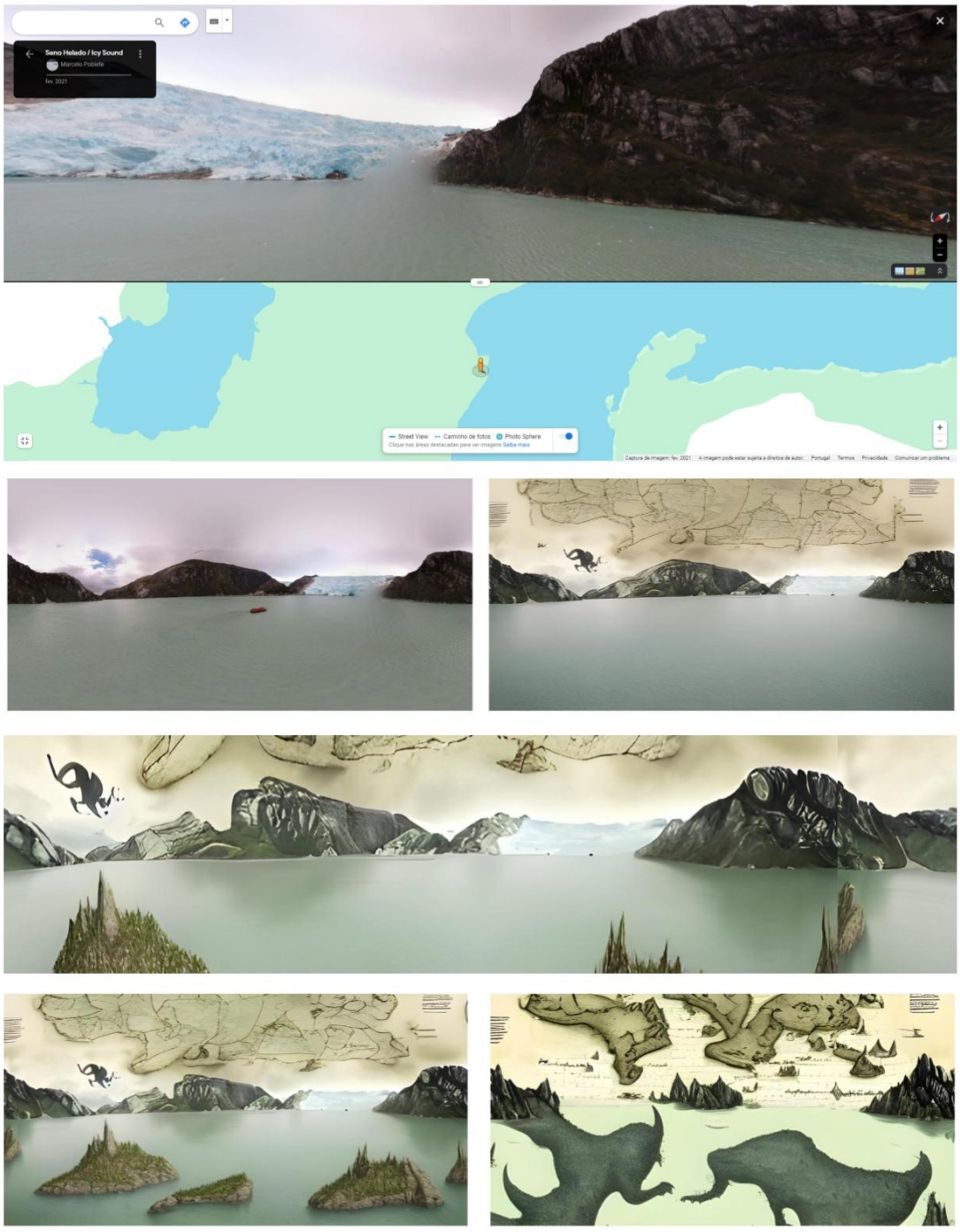


Figure 7.21: Maps are created by searching a location in Google Maps and modify by the text transcribed and the emotional predictions

De vuelta a Casa

Tras la liturgia,
las hermanas limpiaron
los cuartos del convento.
Bromearon sobre un rumor
que poco entendí
y que mal recuerdo.

En la estación de trenes,
alguien habló sobre galerías virtuales.
Yo, en tanto,
mostré entusiasmo por la poesía,
los códigos y el canto.

Es que estoy interesado
en la ambigüedad oral,
o sea:
la discrepancia que se produce
entre lo que se dice
y la forma en que se frasea.

La prosodia del sentido figurado.

Me senté en un parque próximo a nada
y tras observar por un tiempo
a una abeja que no aterrizaba,
se me acercaron dos vagabundos.

Uno de ellos,
peregrino de otro mundo,
dijo algo en italiano,

quizás dirigido a:
dos palomas desplumadas,
a mí,
o a algún otro patibularion
penándonos entre-oculto.

La abeja parecía estar suspendida en el aire.

¿Qué estará haciendo?

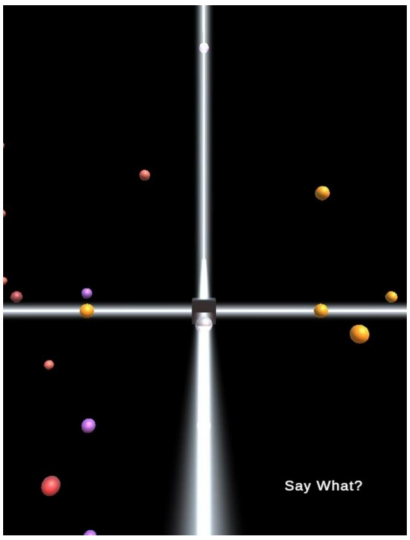
A veces siento
una desbordada emoción
que me angustia
al no poder verbalizarla.

¿Qué estará haciendo?
¿Qué estará haciendo?

Quedamos solo yo y aquel insecto.
Por instantes nos pensamos mutuamente.
Yo queriendo devenir abeja,
Ella deseando ser viento.

Tuve la intención de preguntar,
pero noté que ya era tarde;
el aroma de las flores,
los laberintos, los panales.

El lunático yo,
de vuelta a casa,
Viviendo a momentos.



Sample	Text	Emotion	Valence	Activation	Dominance
Ses03M_impro02_M032	What?	sur	1.5000	3.5000	4.0000
Ses03M_script03_2_F014	What?	ang	1.5000	3.0000	4.5000
Ses05M_impro03_F000	what?	neu	4.0000	2.6667	2.0000
Ses05M_script01_3_M004	What?	exc	2.5000	3.0000	3.0000
Ses03M_script02_2_F010	What?	fru	2.0000	3.0000	2.6667
Ses03M_impro03_F001	What?	exc	4.0000	2.5000	3.0000
Ses03M_impro03_F028	What?	sur	2.5000	2.0000	3.0000
Ses03M_impro03_F029	What?	sur	3.0000	4.5000	4.0000
Ses05M_impro02_F001	What?	sur	2.0000	2.6667	2.0000
Ses05M_impro02_F008	What?	sur	1.3333	3.3333	1.6667
Ses03M_script01_3_M004	What?	neu	3.0000	2.5000	3.5000
Ses01M_script01_1_F010	What?	sur	2.3333	3.0000	3.6667
Ses05F_impro07_M000	what	exc	4.0000	3.0000	3.5000
Ses05M_script03_2_F014	What?	fru	1.5000	3.5000	3.5000
Ses01M_impro03_F000	What?	exc	3.5000	3.0000	2.5000
Ses02F_script01_1_F009	What?	fru	2.5000	2.5000	2.5000
Ses01F_script01_1_F007	What?	ang	2.6667	3.6667	3.6667
Ses05M_impro06_F001	What?	sad	2.0000	3.0000	1.5000
Ses05M_script01_1_F008	What?	fru	2.5000	2.5000	2.5000
Ses04M_script03_2_F019	What?	ang	1.5000	4.0000	3.5000

Figure 7.22: Distribution of "what" utterances in VAD space. Each sphere corresponds to a sample listed in the accompanying table, which details its emotional label and VAD values. The dispersion illustrates how identical lexical content gives rise to distinct affective states through prosodic variation.

7.5 Iteration 7: The What Space

Building on the findings presented in Chapter 5, we became particularly interested in the extreme case of semantic ambiguity represented by short utterances. Among these, the word *what* emerged as the most acoustically variable token in the IEMOCAP dataset, exhibiting a wide distribution of emotional labels despite its minimal semantic content. We developed a series of creative, data-driven audio-visualization strategies aimed at revealing the emotional ambiguity embedded in the utterance *what*. Each strategy explores a different representational paradigm, translating prosodic variation into spatial, visual, and interactive forms. While each system produces an independent output, the overall process unfolded sequentially, with each iteration building conceptually and technically upon the previous one.

Say What? — Interactive Visualization in VAD Space

We first designed an interactive 3D visualization in Unity3D to explore the emotional distribution of the word *what* within the Valence–Arousal–Dominance (VAD) space. Each utterance is represented as a sphere positioned according to its predicted VAD coordinates and color-coded based on its categorical emotion label (see Figure 7.22). Users can orbit a custom reference frame, enabling an interactive exploration of how identical lexical content disperses across affective space. This visualization makes perceptible that prosody alone can generate a multidimensional emotional topology from a single word.[19]

19. Interactive demo available at: https://jfforero.github.io/TheWhatSpace/

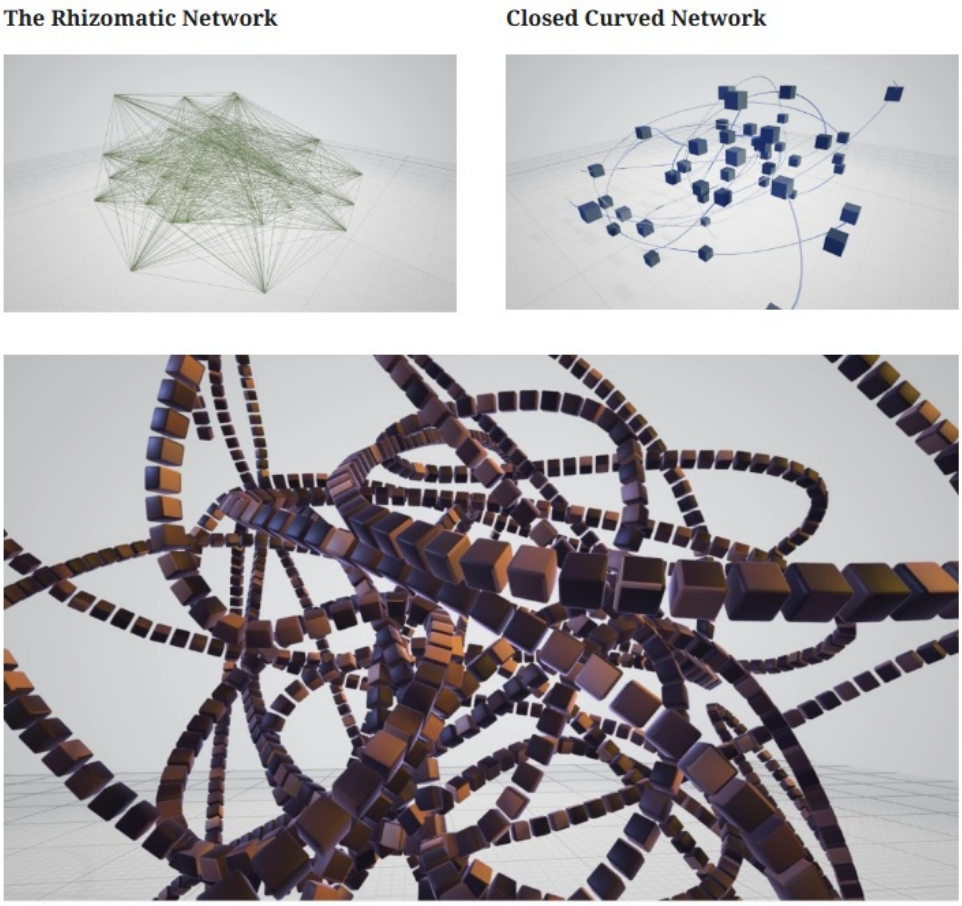


Figure 7.23: Network-based representations of the "What" space. (Top-left) Rhizomatic network showing a fully interconnected, non-hierarchical structure of utterances. (Top-right) Closed sequential network linking samples in a continuous loop. (Bottom) Continuous curved trajectories composed of discrete nodes, suggesting an affective journey through prosodic variation.

The 3D "What" Network — Rhizomatic and Sequential Structures

To further investigate the relational structure of emotional variability, we visualized network topologies using Blender 3D.[20] 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 Figure 7.23).[21]

20. 3D modeling and visualization were implemented using Blender, an open-source software for 3D creation. https://www.blender.org/
21. Interactive 3D model available at: https://skfb.ly/pqRTR (last accessed: March 20, 2026).

Figure 7.24: Emotion-conditioned texture generation from vocal input. Each row corresponds to a predicted emotional category (e.g., neutral, calm, happy, surprised, sad, fearful, angry, disgust), with multiple samples illustrating how predefined prompts translate prosodic features into variations in color, geometry, and illumination.

Speech Emotion to Texture — Real-Time Emotion-to-Image Translation

We developed an other Gradio-based web application, hosted on Hugging Face Spaces[22], that translates vocal emotion into abstract visual textures. The system was based in the architecture introduced in Iteration 6, combining speech-to-text transcription, acoustic emotion recognition and text-to-image synthesis. The interface allows users to record and submit instances of the word "what," encouraging experimentation with different prosodic variations. From the audio input, the system extracts MFCC features to predict an emotional label, while simultaneously transcribing the utterance. Each predicted emotion is mapped to a predefined textual prompt,as shown in table 7.1, which guides the generation of a 512 $\times$ 512 pixel texture (see Figure 7.24).

The "What" Verse — 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 a Mozilla Hubs–based 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 a sequential network configuration (see Figure 7.23). 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 (see Figure 7.25).

22. Huggingface Space Project available at: https://huggingface.co/spaces/jfforero/LoopArtCritique (last accessed: March 20, 2026).

Emotion	Text-to-Image Prompt
Neutral	Generate a geometric texture with neutral colors, balanced illumination, and simple shapes.
Calm	Generate a geometric texture with soft colors and tranquil illumination, using round and calming shapes.
Happy	Generate a geometric texture with vibrant colors and bright, sunny illumination with simple, round shapes.
Sad	Generate a geometric texture with muted colors, somber illumination, and dark and gloomy shapes.
Angry	Create a geometric texture with bold, dark colors, intense illumination, and sharp, irregular shapes.
Fearful	Generate a geometric texture using dark, muted colors with harsh, dim illumination and irregular shapes.
Disgust	Generate a geometric texture with murky, sickly colors, distorted illumination effects, and irregular shapes.
Surprised	Generate a geometric texture with vibrant electric and striking colors, using high-contrast lighting and sharp, angular shapes.

Table 7.1: Emotion-conditioned prompts used to guide text-to-image generation. Each prompt encodes visual attributes such as color, illumination, and geometric structure associated with the predicted emotional state.

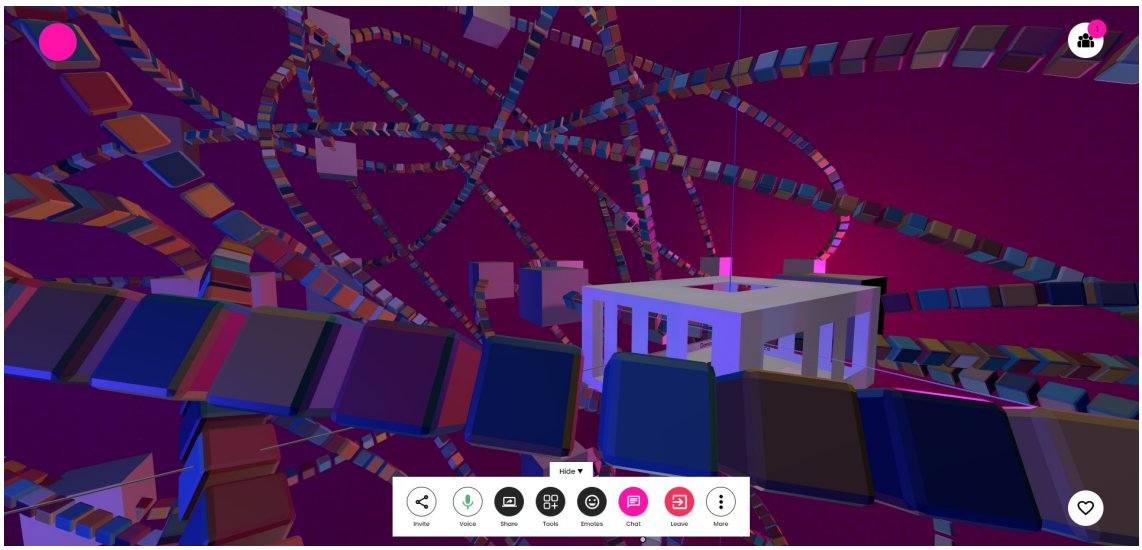


Figure 7.25: Immersive VR environment of the What Space on Onland.io. The scene presents a dense network of interconnected nodes arranged along looping trajectories, where each element corresponds to a recorded utterance of the word “what”. Spatialized structures and chromatic variations reflect affective differences derived from prosodic analysis, allowing users to navigate an emotionally distributed topology through first-person exploration.

7.6 Iteration 8: The Ironic Machine[s]

For Iteration 8, we developed a series of affective virtual environments as a framework for constructing and testing virtual ironic spaces. This iteration builds directly on the data-driven feature analysis and prompt engineering strategy introduced in Chapter 6, particularly the audiovisual feature profiling described in Sections 6.1 and 6.2, and the three-level prompting framework proposed in Section 6.3. The central hypothesis guiding this iteration is that irony can emerge from a structured divergence between expressive modalities. Specifically, we explore irony as a form of cross-modal incongruence between visual and musical emotional cues, where the affective content of one modality contradicts or destabilizes the other.

To operationalize this idea, we employed the statistically significant audiovisual features identified in Chapter 6. From the musical domain, features such as dissonance, harmonic entropy, spectral centroid, and loudness were used to characterize emotional states. From the visual domain, descriptors including brightness, saturation, entropy, edge density, and textural complexity were used to define affective image profiles.

We focused on two opposing emotional categories—*Joy/Happy* (positive valence) and *Anger* (negative valence)—as representative poles for constructing ironic contrasts. These feature profiles were translated into mid-level semantic prompts following the framework defined in Section 6.3, enabling controlled generation of both visual textures and musical audio. Within this framework, irony is formalized as a mismatch between modalities. We therefore defined four possible audiovisual configurations resulting from the combination of visual and musical emotional profiles (see Table 7.2).

Visual Profile	Musical Profile	Audiovisual Space
Anger	Happy	Kind Irony
Anger	Anger	Sincere Anger
Joy	Anger	Sarcasm
Joy	Happy	Sincere Happiness

Table 7.2: Audiovisual profile combinations used to generate the Ironic Machine[s].

In this formulation, *Sincere* environments correspond to aligned emotional modalities, while *Sarcasm* and *Kind Irony* emerge from cross-modal opposition. This structure provides a computational analogue to verbal irony, where meaning arises, as seen in chapter 5, through contradiction between semantic and acoustic layers.

For visual generation, we used the *DeepAI* API to produce abstract texture images conditioned on the feature-derived prompts[23]. For each emotional profile, nine variations were generated per prompt (see Fig. 7.26), allowing exploration of variability within each affective configuration. For the musical component, we generated nine 30-second audio clips per emotional category using **MusicGen**[24].

23. Obtained from https://deepai.org/machine-learning-model/text2img
24. Obtained from https://colab.research.google.com/github/sanchit-gandhi/notebooks/blob/main/MusicGen.ipynb

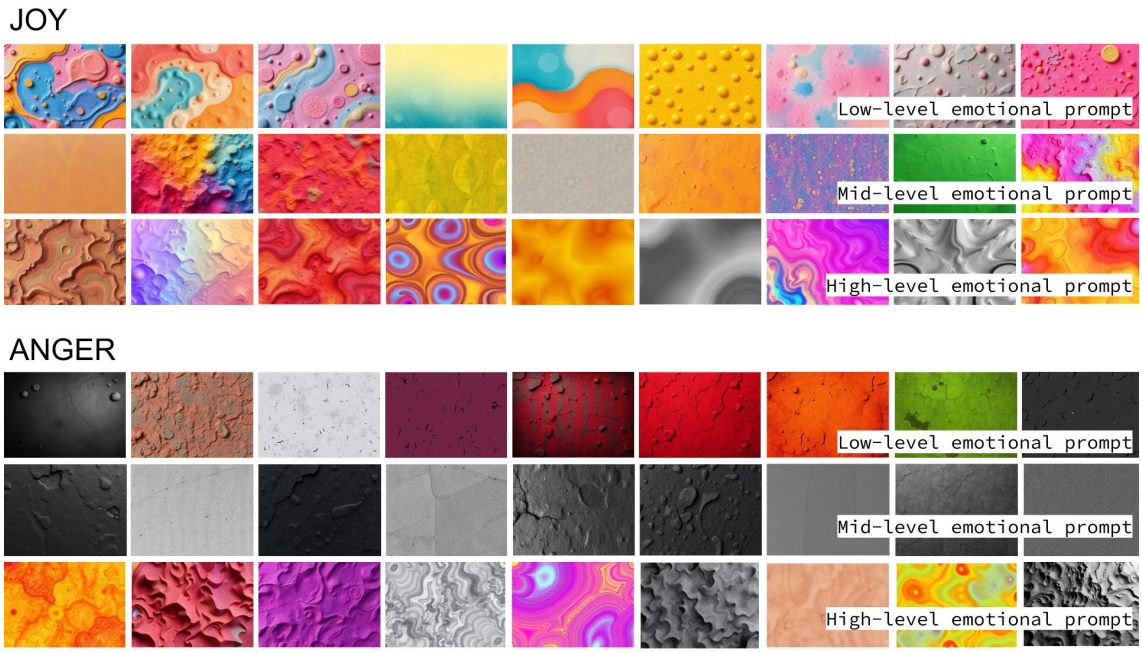


Figure 7.26: Examples of AI-generated texture images for two contrasting emotional categories—*Joy* (top) and *Anger* (bottom)—across three levels of prompt specificity. Each row corresponds to a different prompting strategy derived from the framework in Section 6.3: low-level (feature-based parametric descriptions), mid-level (semantic descriptors grounded in statistically significant audiovisual features), and high-level (abstract emotional labels). The results illustrate how increasing levels of abstraction influence visual appearance, with low-level prompts producing more controlled and homogeneous textures, and high-level prompts yielding more diverse and stylistically varied outputs.

The resulting audiovisual assets were integrated into an interactive web-based application developed in p5.js. This interface allows users to explore the four configurations dynamically, enabling direct comparison between aligned and mismatched emotional conditions (see Figure 7.27).

The prompts used for image generation were explicitly conditioned on the statistical distributions identified in the visual analysis:

- **Happy:** Bright, highly saturated color distributions, high entropy, strong symmetry, dense edge structures, and fine-grained textures, reflecting high visual complexity and positive valence.
- **Anger:** Darker tonal ranges, reduced brightness and color variance, lower entropy, asymmetrical composition, and coarser textures, reflecting negative valence and structural tension.

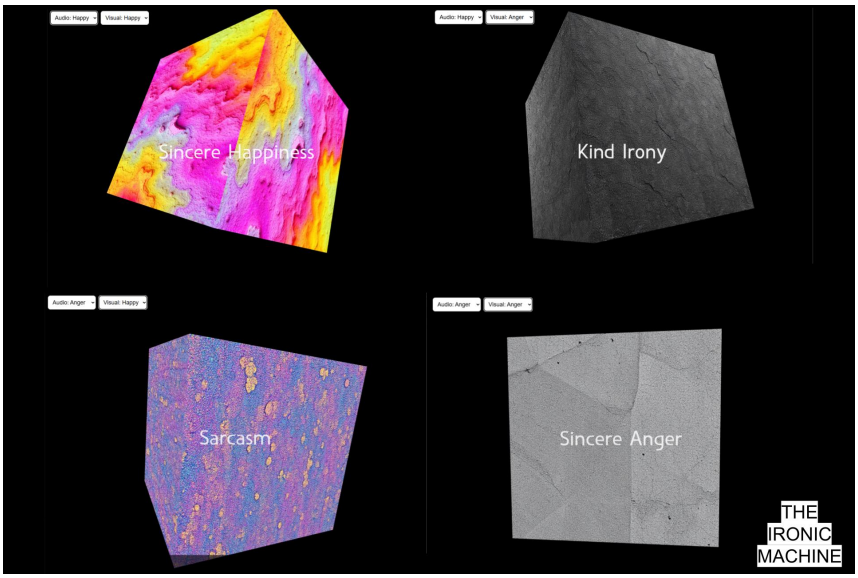


Figure 7.27: Interface of the *Ironic Machine* implemented in p5.js, displaying the four audiovisual configurations derived from cross-modal emotional combinations. Each panel represents a distinct condition: *Sincere Happiness* (top-left), *Kind Irony* (top-right), *Sarcasm* (bottom-left), and *Sincere Anger* (bottom-right). The system allows real-time selection of visual and musical emotional profiles, enabling the exploration of alignment and divergence between modalities as a mechanism for producing ironic affective environments.

Finally, as shown in figure 7.27, a graphical user interface was developed to allow users to select emotional configurations for both visual and musical generation based on mid-level semantic descriptors[25]. Through this interface, the system functions as an exploratory instrument for navigating affective incongruence, enabling users to experience how emotional meaning shifts when modalities align or diverge.

7.7 Iteration 9: Bello

Iteration 9 culminates in a speculative reflection on the multidimensional nature of irony, articulated through the cultural expression *"más perdido que el Teniente Bello"* (more lost than Lieutenant Bello). This phrase refers to the disappearance of Teniente Alejandro Bello Silva during a flight in 1914 in Chile, an event that remains unresolved and has since entered popular language as a metaphor for disorientation and loss. The irony embedded in this expression operates across multiple layers and can be interpreted as a form of *kind irony*, invoking a tragic historical figure through a seemingly light or colloquial (joyful) remark.

Building on the architecture of previous iterations, we developed a new instance of the Affective Virtual Environment (AVE) system, deployed on Hugging Face, designed to generate immersive 360° audiovisual environments (equirectangular imagery combined with generative sound)

25. Available at https://jfforero.github.io/TheIronicMachine/index.html

[26]. In this iteration, participants are invited to speculate on the possible fate of Teniente Bello by recording spoken utterances. These utterances serve as the primary input for a multimodal pipeline that integrates speech-to-text transcription, acoustic emotion recognition, sentiment analysis, and generative audiovisual synthesis. The system extends the pipeline introduced in Iteration 6. Audio input is first segmented into temporal chunks, enabling localized analysis of speech. For each segment, transcription is performed using the *Whisper* model, while emotional prediction is derived from MFCC-based features processed through a pre-trained LSTM network. In parallel, semantic sentiment is estimated from the transcribed text. These three layers—acoustic emotion, semantic sentiment, and linguistic content—jointly inform the generation process.

Visual output consists of high-resolution equirectangular panoramas generated through text-to-image synthesis, guided by structured prompts that encode both semantic content and affective features. In our implementation, prompt generation follows a rule-based strategy in which sentiment polarity modulates visual attributes such as brightness, color distribution, texture frequency, spatial complexity, and edge density. Positive sentiment leads to high-frequency textures, bright tonal ranges, and increased structural complexity, whereas negative sentiment produces darker, lower-frequency, and more spatially reduced compositions. Neutral sentiment results in balanced configurations across these parameters. This mapping translates affective inference into controllable visual descriptors (see Table 7.3).

Sentiment	**Image Generation Prompt Structure**
Positive	Generate an equirectangular 360° panoramic graphite sketch with detailed pencil texture and faint neon glows, using cinematic lighting based on the [**transcribed text**]. Emphasize high brightness, high color variance, dominant high RGB values, and low histogram frequency in bright bins. Apply high-frequency textures with strong gradients, high local contrast, high spatial complexity, strong symmetry, and high edge density, resulting in rich visual structure and intense variation.
Negative	Generate an equirectangular 360° panoramic graphite sketch with detailed pencil texture and faint neon glows, using cinematic lighting based on the [**transcribed text**]. Emphasize low brightness, low color variance, dominant low RGB values, and high histogram frequency in dark bins. Apply low-frequency textures with weak gradients, low local contrast, low spatial complexity, reduced symmetry, and low edge density, resulting in coarse structure and limited variation.
Neutral	Generate an equirectangular 360° panoramic graphite sketch with detailed pencil texture and faint neon glows, using cinematic lighting based on the [**transcribed text**]. Maintain balanced histogram distribution, mid-range RGB values, and moderate brightness and color variance. Apply medium-frequency textures with moderate gradients, average contrast, balanced symmetry, and medium edge density, resulting in naturalistic and stable visual structure.

Table 7.3: Sentiment-conditioned prompt generation for equirectangular image synthesis. Visual attributes are modulated according to sentiment polarity derived from transcribed speech. Note that the [**transcribed text**] is insert in the prompt

In parallel, a generative music model (*MusicGen*) produces short audio compositions conditioned on acoustic emotion predictions. Each emotional category is associated with a predefined

26. Available at https://huggingface.co/spaces/jfforero/Bello.

prompt structure that modulates rhythmic density, spectral characteristics, harmonic motion, and dynamic range. For instance, positive emotions such as happiness generate rhythmically dense, tonally stable, and bright timbral outputs, whereas negative emotions such as anger or fear introduce dissonance, instability, and irregular temporal structures. This approach enables the systematic translation of vocal affect into musical behavior (see Table 7.4).

Emotion	**Music Generation Prompt Structure**
Neutral	Balanced orchestral soundtrack with steady tempo, even rhythmic density, low dissonance, moderate pitch clarity, stable loudness, and slow harmonic motion, producing an ambient and unobtrusive atmosphere.
Calm	Slow orchestral soundtrack with sparse rhythm, warm timbre, minimal dissonance, gentle pitch presence, restrained dynamics, and stable tonality, emphasizing tranquility and sustained harmonic flow.
Happy	Energetic orchestral soundtrack with lively rhythm, bright timbre, clear melodic focus, dynamic expressiveness, and stable tonal grounding, emphasizing playful and uplifting harmonic progressions.
Sad	Slow and sparse orchestral soundtrack with dark timbre, subdued pitch clarity, restrained dynamics, minimal harmonic change, and minor tonal coloration, emphasizing fragility and emotional depth.
Angry	Fast and dense orchestral soundtrack with sharp timbre, high dissonance, strong rhythmic attack, high dynamics, and unstable tonal grounding, emphasizing tension and aggressive articulation.
Fearful	Unstable orchestral soundtrack with fluctuating rhythm, variable timbre, shifting dissonance, blurred pitch, and volatile dynamics, emphasizing suspense and unpredictability.
Disgust	Irregular orchestral soundtrack with rough timbre, abrasive dissonance, unstable spectral character, weak pitch focus, and uneven dynamics, emphasizing distortion and tonal ambiguity.
Surprised	Dynamic orchestral soundtrack with sudden rhythmic variation, sharp contrasts, expressive pitch clarity, and rapid harmonic shifts, emphasizing unpredictability and expressive transitions.

Table 7.4: Emotion-conditioned prompt generation for music synthesis using MusicGen. Acoustic emotion predictions modulate temporal, spectral, and harmonic properties of the generated audio.

The result is a sequence of audiovisual fragments, each corresponding to a segment of the user's speech. The system ultimately compiles these fragments into an interactive, HTML-based 360° environment, where users can navigate between segments and experience an evolving affective landscape. Although the system has been trained in English, it can be applied to any language through *Whisper*-based translation and transcription. In this iteration, the system is fed with textual material derived from the poem included in the song *El Teniente Bello* by Mauricio Redolés[27]. The generated audiovisual sequences correspond to 12-second segments of processed speech and can be accessed through the project's interactive web interface[28] (see Figure 7.28).

The material produced by the AVE system was collected and subsequently used to create audiovisual content, including a mapping projection presented at the Algarve Design Meeting 2025 (see Figure 7.3).

27. Mauricio Redolés, "El Teniente Bello," in *Bello Barrio*, 1987.
28. https://jfforero.github.io/El-Teniente-Bello_AVEbyRedoles/index.html

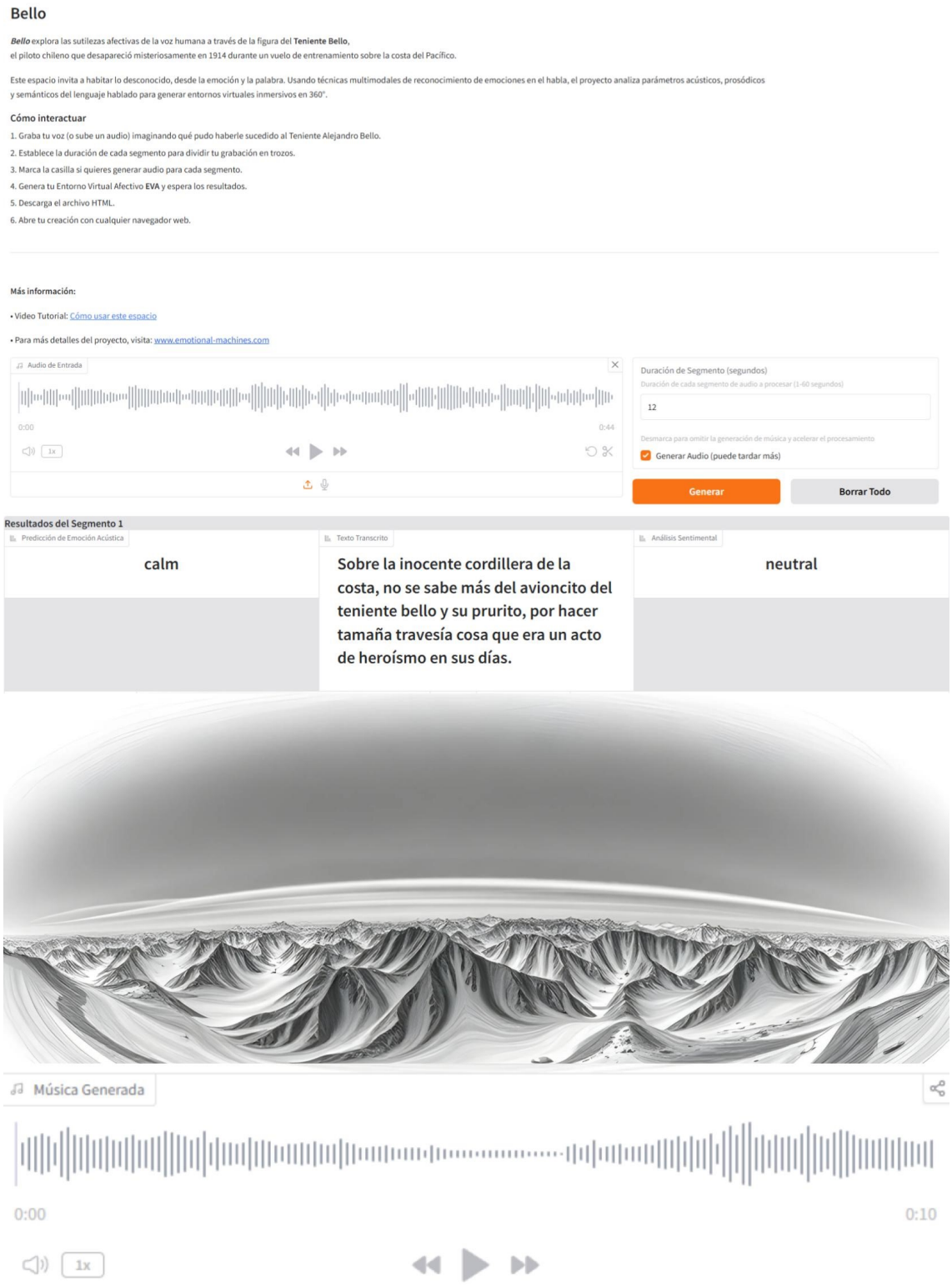


Figure 7.28: Excerpt from the poem *El Teniente Bello* by Mauricio Redolés, originally featured in the album *Bello Barrio* (1987). The poem reinterprets the historical disappearance of Teniente Bello through a political and poetic lens, linking themes of memory, loss, and national identity. In this work, the text is incorporated as a semantic and narrative substrate within the generative system, connecting literary discourse with speech-driven affective analysis and audiovisual synthesis.

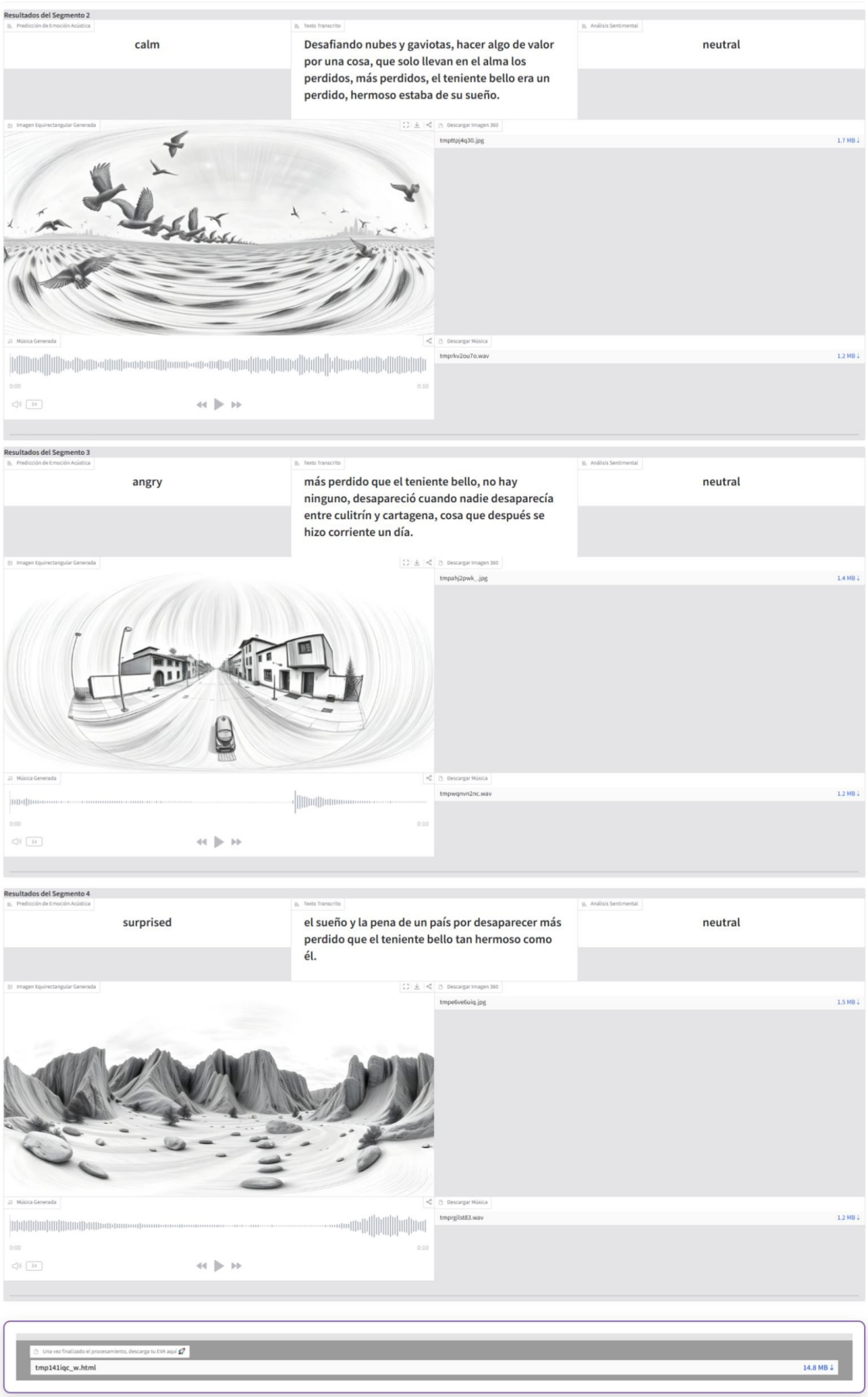


Figure 7.28: (continued)

Summary

Across the nine iterations presented in this chapter, the research progressively evolves from a functional integration of speech-driven interaction into a broader investigation of affect, ambiguity, and multimodal meaning-making within computational environments. What begins in Iterations 1 and 2 as a system for mapping speech to sound and object manipulation gradually unfolds into a complex assemblage of interconnected systems in which language, emotion, image, and music operate as co-constitutive layers of experience.

A central trajectory throughout these developments is the increasing displacement of semantic determinacy in favor of affective modulation. Early implementations relied primarily on lexical sentiment analysis, where meaning was inferred from textual content. However, subsequent iterations introduced acoustic emotion recognition as a parallel—and often conflicting—channel of interpretation. This shift enabled the emergence of ambiguity as a productive condition rather than a limitation, particularly evident in Iteration 7 (*The What Space*) and Iteration 8 (*The Ironic Machine[s]*), where divergence between modalities becomes the primary generator of meaning.

The notion of *[no]-return* and *[no]-belonging*, introduced at the beginning of this chapter, finds its operational realization in the way environments are constructed and experienced. Projects such as *Abandoned Memories*, *Veintitantos Vientos*, and *En train d'oublier* explore displacement through geographically anchored yet affectively transformed landscapes, while *Terra Australis Ignota* extends this logic into a speculative cartography where territory is reimagined through emotional inference. In these works, belonging is not tied to physical location, but emerges as a transient relation between voice, memory, and generated space.

The final iterations consolidate these explorations into a formalization of irony as a computational and experiential phenomenon. By defining irony as a structured divergence between modalities—whether between semantic and acoustic layers or between visual and musical domains—the research introduces a novel framework for understanding affective incongruence in artificial systems. The *Ironic Machine[s]* and *Bello* exemplify this approach, transforming irony from a linguistic figure into an operational principle embedded within generative pipelines.

Importantly, the artifacts produced throughout this process do not function solely as technical demonstrations, but as autonomous creative works that inhabit the intersection of media art, affective computing, and digital poetics. Their presentation across diverse contexts—ranging from academic conferences to immersive exhibitions and virtual platforms—highlights the dual nature of the research as both analytical and experiential.

In this sense, the chapter articulates a practice-based methodology in which artistic production operates as a mode of inquiry. Each iteration constitutes not only a technical advancement but also a conceptual probe, testing how computational systems can engage with ambiguity, emotion, and the instability of meaning. The resulting body of work suggests that affective virtual environments can serve as experimental infrastructures for exploring the limits of language and the conditions under which meaning emerges, fractures, and reconfigures.

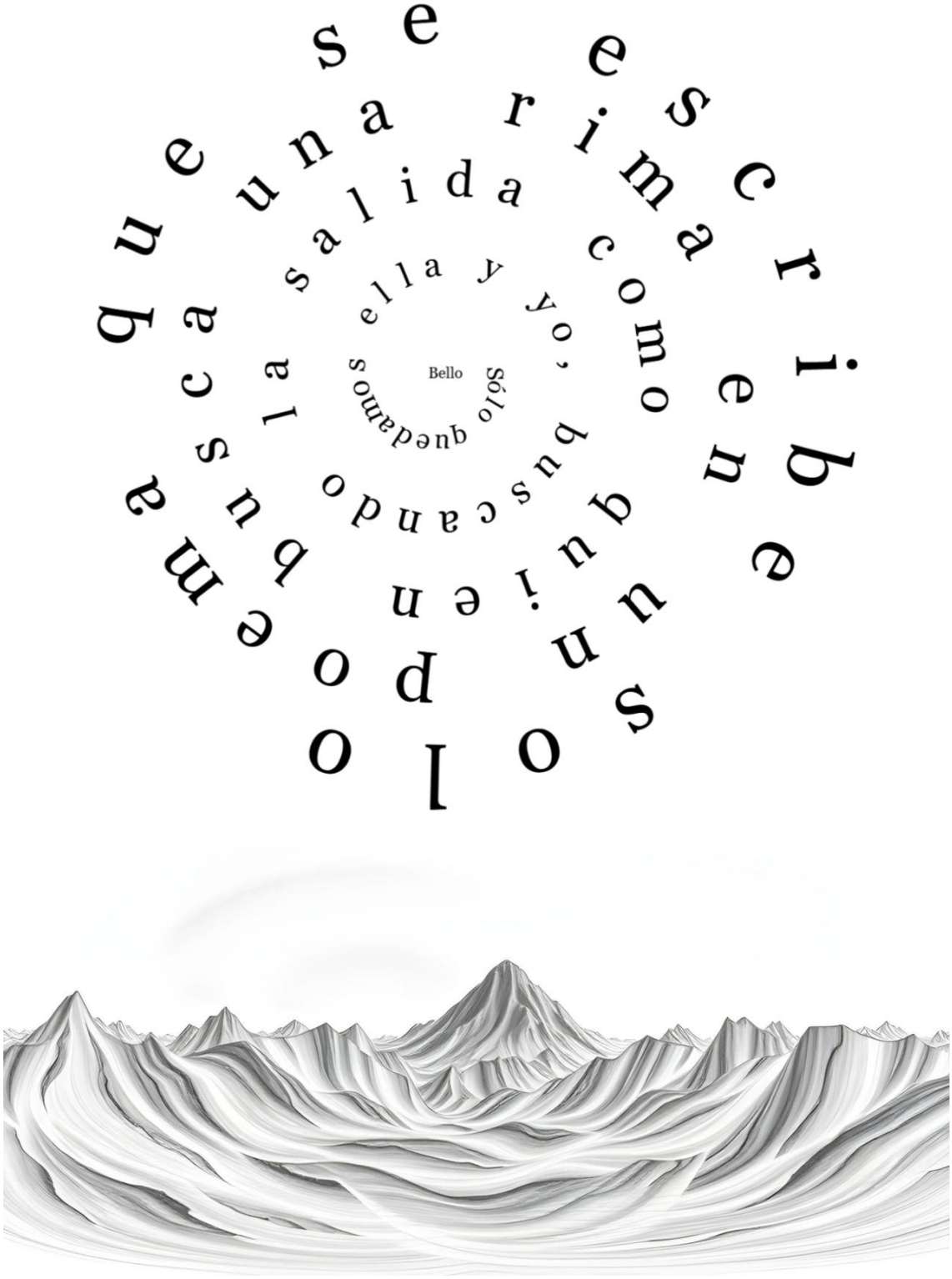


Figure 7.29: Visual representation inspired by the figure of Teniente Bello, integrating poetic structure and generative imagery.

Chapter 8

Conclusion

This research proposed and implemented a methodological framework for the analysis of prosodic irony and its translation into Affective Virtual Environments. Situated at the intersection of media art, affective computing, and natural language processing, the study investigated how rhetorical phenomena such as kind irony and sarcasm emerge from systematic divergences between semantic and acoustic emotional predictions.

These divergences were formalized as a measurable multimodal phenomenon, integrating sentiment analysis, speech emotion recognition, perceptual evaluation, and statistical inference. Through this approach, the research identified prosodic features that significantly differentiate sincere and ironic speech, demonstrating that irony can be understood as an empirically observable interaction between linguistic meaning and vocal affect.

A secondary contribution of this work was the exploration of how these analytically derived affective descriptors can be operationalized within generative systems. A three-level prompt engineering framework was developed to map semantic and acoustic emotional predictions onto audio-visual features extracted from musical and visual datasets. This framework enabled the translation of multimodal affective analysis into strategies for constructing Affective Virtual Environments, where emotional expression is shaped through structured computational design.

The research was developed through nine technological and artistic iterations, conceived as autonomous yet interconnected experimental platforms. These works functioned as sites for investigating emotional ambiguity, perceptual disorientation, and the instability of meaning within computational systems. Across these iterations, irony was treated not as a fixed semantic category, but as a dynamic cross-modal phenomenon emerging from controlled divergences between modalities.

Collectively, this thesis proposes both a methodological model for the analysis of prosodic irony and a speculative framework for affect-driven generative environments. It positions divergence between semantic and acoustic representations as a central operational principle in multimodal systems, and demonstrates how practice-based research can function as a rigorous mode of inquiry into the relationship between human emotion and machine-mediated expression.

8.1 Dissertation Contributions

The Bimodal Speech Analysis

The research established a dual-stream emotion recognition system combining semantic sentiment analysis with acoustic speech emotion recognition. The semantic channel employed three classifiers of increasing representational capacity: Rocchio classifier achieving 0.595 accuracy, LSTM with GloVe embeddings reaching 0.694 accuracy, and BERT based models attaining 0.728 accuracy. These results demonstrate that contextualized transformer representations most effectively capture pragmatic emotional cues compared to prototype-based or sequential architectures. The acoustic channel extracted emotional content through multilayer perceptron, convolutional neural network, and CNN+LSTM architectures trained on eGeMAPS features. The CNN+LSTM hybrid achieved 0.6842 overall accuracy with balanced per-class performance, particularly strengthening neutral classification (F1 = 0.71) and happy classification (F1 = 0.77).

The adequacy of each modality was defined not by proximity to isolated benchmarks, but by capacity to sustain coherent contrasts within a bimodal system. This criterion enabled reliable comparison between semantic and acoustic predictions, establishing the foundational architecture for irony detection. We acknowledge that the semantic sentiment models achieved accuracies below current state-of-the-art results exceeding 0.80, a gap primarily attributable to the single-modal text-only evaluation protocol and deliberate exclusion of ensemble techniques or data augmentation strategies that fall outside the scope of controlled architectural comparison. Similarly, acoustic emotion recognition on fixed-length functional descriptors (71.18% for MLP, 68.42% for CNN+LSTM) reflects a deliberate choice prioritizing interpretability and per-class robustness, particularly for neutral and happy categories critical to downstream irony detection, over maximizing raw accuracy through spectrogram-based deep learning approaches.

Prosodic Characterization of Irony

Multimodal analysis of the IEMOCAP corpus identified 51 kind-irony utterances (37 female, 14 male) and 28 sarcasm utterances (6 female, 22 male), contrasted against matched sincere controls comprising 51 sincere anger utterances (31 female, 20 male) and 28 sincere happy utterances (14 female, 14 male). We acknowledge considerable gender imbalance in the sarcasm condition, which limits the generalizability of gender-stratified findings for this irony type and should be interpreted as preliminary until replication with larger, more balanced samples.

Kind irony was characterized by an attenuation strategy: pitch reduction of approximately 4.57 semitones in male speakers ($p < .01$) and 2.38 semitones in female speakers ($p < .001$), intensity reduction observed in both genders ($p < .001$), and increased voice clarity measured by harmonics-to-noise ratio in female speakers (female HNR: +13.05%, $p < .01$). Temporal structure remained largely preserved between kind-irony and sincere anger conditions. Sarcasm operated through an amplification strategy characterized by pitch elevation of 3.26 semitones in male speakers and 5.10 semitones in female speakers ($p < .05$ for both genders), substantial loudness increases of

8.21 decibels in males and 9.85 decibels in females ($p < .01$), and accelerated speech rate in male speakers ($p < .01$).

Principal Component Analysis revealed that prosodic features organize into two partially dissociated dimensions: PC1 capturing 42 to 50 percent of variance in energetic and temporal properties, and PC2 capturing 26 to 27 percent of variance in spectral properties. Analysis suggests that kind irony in this corpus is characterized by variations primarily along PC2, while sarcasm engages both dimensions. Perceptual validation confirmed these distinctions: kind-ironic stimuli received mean irony ratings of 3.31 versus 2.06 for sincere anger (Mann-Whitney $U = 149.0$, $p = .0010$, Cliff's $\delta = 0.76$); sarcastic stimuli received 3.80 versus 2.45 for sincere happiness ($U = 146.5$, $p = .0016$, $\delta = 0.73$).

The observed pattern of pitch reduction without corresponding increase in pitch variability suggests that kind irony does not operate through exaggerated intonational modulation, but rather through a global downward shift in pitch register. The pragmatic effect appears to be one of affective detachment: the speaker attenuates the acoustic markers of genuine emotional engagement while maintaining lexical content that implies it. For female speakers, the additional increase in harmonics-to-noise ratio suggests deployment of a broader acoustic repertoire in signaling kind irony, recruiting phonatory quality alongside pitch and intensity, whereas male ironic expression relies more selectively on global shifts in energetic and spectral dimensions. This gender-differentiated finding requires confirmation with larger and more balanced gender samples before being treated as a robust sex-based distinction.

Affective Virtual Environment Generation

Statistical analysis of audiovisual features identified eleven musical features (HPCP entropy mean: $\eta^2 = 0.791$; spectral centroid: $\eta^2 = 0.732$) and 54 visual features (color brightness: $\eta^2 = 0.109$) as significantly discriminating emotional categories according to Benjamini-Hochberg false discovery rate correction. These empirically derived descriptors grounded a three-level prompting framework: high-level emotional labels providing intuitive but ambiguous specifications; mid-level feature-derived descriptions anchoring specifications to statistically significant effects; and low-level parametric prompts specifying approximate feature values derived from dataset distributions.

Perceptual evaluation comparing high-level and mid-level prompts across eight audiovisual environments revealed that mid-level prompts produced lower audiovisual mismatch ratings (2.24 versus 3.00, representing a 25 percent improvement, $p = 0.07$). This directional advantage did not reach conventional statistical significance thresholds, likely due to limited sample size (30 ratings across 8 environments), and should be interpreted as exploratory evidence rather than conclusive demonstration of prompt strategy superiority. The non-significant Mann-Whitney U test comparing ironic versus sincere audiovisual configurations ($U = 93$, $p = 0.39$) similarly suggests that while environments expressing cross-modal emotional opposition tended toward slightly higher mismatch scores consistent with the conceptual analogy to verbal irony, this pattern was not statistically robust given available data.

A fundamental limitation emerged during generative implementation: contemporary generative models respond probabilistically rather than deterministically to feature specifications, meaning identical prompts produce variable outputs and no mechanism exists for ensuring consistent emotional signatures across generations. The interpretability-to-latency chasm, defined as the persistent gap between human-readable feature specifications and opaque latent generative processes, prevents deterministic affective control. For creative and exploratory purposes, this probabilistic behavior serves artistic intent by introducing generative variability conducive to interpretation and meaning-making.

Creative Practice as Methodology

Nine technological and artistic iterations functioned as experimental platforms for examining emotional ambiguity, disorientation, and meaning instability within computational systems. Rather than resolving ambiguity, these works sustained it as a productive space where affect, technology, and meaning negotiate one another. Early iterations established foundational integration of real-time speech processing with audiovisual generation; subsequent works explored displacement through geographically anchored yet affectively transformed landscapes; final iterations operationalized irony as structured divergence between modalities.

8.2 Research Questions Addressed

RQ1: Can multimodal speech emotion recognition detect and characterize prosodic verbal irony? Yes, within the constraints of acted, laboratory speech. The bimodal framework successfully detects irony through contradictions between lexical polarity and acoustic emotional predictions. Independent calibration of semantic (BERT: 72.8% accuracy on DynaSent lexical polarity) and acoustic (CNN+LSTM: 68.42% accuracy on RAVDESS emotions) channels ensured balanced but distinct predictions. Cross-modal filtering identified 51 kind-irony and 28 sarcasm utterances from the IEMOCAP database exhibiting systematic lexical-prosodic divergence. Perceptual validation with six native English speakers confirmed human recognition of these bimodal configurations (kind irony: Cliff's $\delta = 0.76$, 95% CI [0.46, 0.99]; sarcasm: $\delta = 0.73$, 95% CI [0.40, 0.98]), establishing that prosodic irony emerges reliably from cross-modal incongruence in controlled settings.

RQ2: Which prosodic features differentiate ironic from sincere speech? Ironic speech employs two prosodically opposed strategies, differentiated by gender and mapped to distinct principal components.

Kind irony operates through **affective downshifting** - a coordinated attenuation across spectral and energetic dimensions while temporal structure is preserved. Specifically: (1) pitch lowering relative to sincere anger baseline (female: $\Delta = -2.38$ semitones, $p = 0.001$; male: $\Delta = -4.57$ semitones, $p = 0.006$); (2) intensity reduction (both genders, $p < 0.001$; female: $\Delta M = -8.18$ dB, male: $\Delta M = -10.72$ dB); (3) increased voice clarity in female speakers only (HNR: $+13.05\%$,

$p = 0.008$, $\eta^2 = 0.076$). Temporal features (speech rate, articulation rate, syllable duration) showed no significant differences ($p > 0.20$), suggesting that kind irony preserves the rhythmic baseline of ordinary interaction despite affective modulation. PCA analysis revealed that pitch-related features load substantially on PC2 (26 to 27% variance), while intensity and temporal features dominate PC1 (42% and higher variance), indicating a dissociated two-tier structure wherein spectral register (F0) functions as the primary carrier of ironic intent.

Sarcasm operates through **expressive amplification** - systematic intensification across multiple acoustic dimensions with pronounced gender asymmetry in temporal engagement. Specifically: (1) pitch elevation (female: $\Delta = +5.10$ semitones, $p = 0.019$; male: $\Delta = +3.26$ semitones, $p = 0.012$); (2) substantial loudness increases (female: $\Delta M = +9.85$ dB, male: $\Delta M = +8.21$ dB, both $p < 0.01$); (3) increased pitch variability (StdvF0 elevation, particularly in females); (4) temporal acceleration *in male speakers only* (articulation rate: $p < 0.01$; syllable duration: $p < 0.01$), indicating that male sarcasm recruits speech motor acceleration as an additional expressive resource. PCA revealed that MeanF0 and IntensityAvg co-dominate PC1 (41 to 50% variance across genders), while temporal features load on higher-order components (PC3 to PC5) for males, suggesting a coordinated expressive activation dimension distinct from the spectral-only modulation of kind irony.

RQ3: How can bimodal emotional representations be mapped into affective virtual environments? Bimodal emotional prediction (lexical polarity plus acoustic emotion) combined with speech transcription can inform semantic prompts for generative audiovisual creation. A three-level prompting framework was developed: (1) high-level (emotional label only, e.g., “angry”); (2) mid-level (feature-derived semantic descriptors grounded in FDR-corrected statistical effects, e.g., “sharp pitch clarity, rapid harmonic changes, high dissonance”); (3) low-level (numerical parameter ranges derived from dataset means, e.g., “dissonance approximately 0.49, spectral centroid approximately 2140 Hz”).

A pilot perceptual evaluation (N equals 30 ratings across 8 audiovisual environments) measured perceived alignment between visual and musical modalities using a five-point mismatch scale. Environments generated with mid-level feature-derived prompts showed numerically lower mismatch scores (M equals 2.24) than those generated with high-level prompts (M equals 3.00), suggesting a 25% improvement in perceived audiovisual coherence. However, this difference approached but did not reach statistical significance (Mann–Whitney $U = 73$, $p = 0.07$), and post-hoc power analysis indicates that the sample size (n equals 4 scenes per condition) is *severely underpowered* to detect meaningful effects.

RQ4: Can audiovisual incongruence function as an analogue of verbal irony? No, or only in a limited and metaphorical sense not supported by the current data. While audiovisual emotional incongruence can be *constructed* using the bimodal logic that characterizes verbal irony (e.g., pairing negative visual content with positive audio), the perceptual salience of this incongruence does not parallel that of verbal irony.

Environments explicitly designed to express cross-modal emotional opposition achieved the

following mean mismatch ratings: (1) kind irony (negative visual plus positive audio): M equals 2.71, SD equals 1.36; (2) sarcasm (positive visual plus negative audio): M equals 2.83, SD equals 1.22. For comparison, sincere conditions: (3) happy (positive plus positive): M equals 2.63, SD equals 1.12; (4) anger (negative plus negative): M equals 2.22, SD equals 1.35.

A Mann–Whitney U test comparing ironic conditions (kind irony plus sarcasm, aggregated) against sincere conditions revealed **no statistically significant difference** (U equals 93, p equals 0.39). In fact, sincere anger achieved the *lowest* mismatch score (most perceived alignment), directly contradicting the hypothesis that audiovisual opposition should be perceptually salient as a form of "irony."

8.3 Limitations and Future Directions

This research establishes irony as a measurable multimodal phenomenon through controlled analysis of acted speech (IEMOCAP) and generative audiovisual environments, but its findings are constrained by several methodological and epistemological limitations that define clear boundary conditions for generalization.

First, the reliance on the IEMOCAP corpus introduces a strong ecological validity constraint, as the dataset consists of professionally acted, highly controlled emotional expressions where irony is acoustically salient and intentionally performed. This limits applicability to spontaneous speech, where ironic meaning is typically more subtle, context-dependent, and distributed across discourse. Future work must therefore validate prosodic irony markers in naturalistic conversational corpora. Additionally, speaker diversity is limited (ten actors), and the sarcasm subset exhibits strong gender imbalance (6 female vs. 22 male samples), restricting the reliability of gender-specific temporal findings and reducing statistical power for subgroup analyses.

Second, the emotional scope is restricted to four high-arousal categories (kind irony, sarcasm, sincere anger, sincere happiness), excluding low-arousal, mixed, and broader irony typologies (e.g., situational or dramatic irony). This narrows the framework to a specific subset of irony characterized by explicit semantic–prosodic opposition.

Third, the semantic channel relies on BERT-based sentiment classification (72.8% accuracy), which provides only coarse polarity labels and cannot capture pragmatic meaning, discourse context, or implicature. This introduces classification noise ($\sim$27%) that may propagate into irony detection and misrepresent semantic–prosodic divergence in context-dependent utterances.

Fourth, the generative audiovisual framework is constrained by model-specific stochasticity and prompt sensitivity. Results from MusicGen and diffusion-based visual models indicate that prompting effectiveness varies across abstraction levels (high-, mid-, low-level), but these effects are not necessarily generalizable across architectures. The inherent stochastic nature of generative models creates an interpretability–control gap between human-readable feature specifications and latent generative processes, limiting deterministic control of affective outputs.

Fifth, perceptual evaluation results are underpowered ($n \approx 30$ ratings across eight stimuli), preventing robust statistical inference regarding audiovisual incongruence. Although trends suggest higher mismatch perception in ironic conditions, these effects are not statistically reliable. The audiovisual irony analogue also shows substantially weaker perceptual salience than prosodic irony, indicating that cross-modal opposition is not yet equivalent in strength to semantic–prosodic incongruence in speech.

Finally, audiovisual feature mappings are derived from specific domains (Pop/Rock music and naturalistic images), limiting generalization to other musical genres, visual styles, and synthetic domains.

Despite these limitations, the framework demonstrates that irony can be operationalized as cross-modal divergence and translated into generative systems. Future research should focus on spontaneous and cross-linguistic datasets, larger and balanced corpora, causal experimental designs of irony production, systematic comparisons across generative architectures, and the development of more deterministic or feature-constrained generation methods. Extending the framework to interactive, narrative-situated environments may further clarify the role of temporal context and pragmatic inference in audiovisual irony perception.

References

Abadi, Martín, et al. 2016. "TensorFlow: A System for Large-Scale Machine Learning." In *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation,* 265–283.

Agostinelli, Andrea, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. 2023. "MusicLM: Generating Music From Text." *arXiv preprint arXiv:2301.11325,* https://arxiv.org/abs/2301.11325.

Akbari, Hassan, Linjie Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. "VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text." In *Advances in Neural Information Processing Systems.*

Akhtar, Muhammad Arsalan. 2021. "HeartBeat: Visualizing Collective Emotion During the COVID-19 Pandemic." Master of Design, Digital Futures, OCAD University. https://openresearch.ocadu.ca/id/eprint/3306/1/AKHTAR_MUHAMMAD%20ARSALAN_2021_MDES_DIGF.pdf.

Allen, James G., and J. R. Gold. 1986. "Psychometric Problems in the Measurement of Mood." *Journal of Research in Personality* 20 (2): 186–199.

Anderson, Page L., Matthew Price, Shannan M. Edwards, Marilyn A. Obasaju, Stefan K. Schmertz, and Elana Zimand. 2013. "Virtual Reality Exposure Therapy for Social Anxiety Disorder: A Randomized Controlled Trial." *Journal of Consulting and Clinical Psychology* 81 (5): 751–760. https://doi.org/10.1037/a0033559.

Anolli, Luigi, Rita Ciceri, and Maria Grazia Infantino. 2000. "Irony as a Game of Implicitness: Acoustic Profiles of Ironic Communication." *Journal of Psycholinguistic Research* 29 (3): 275–311. https://doi.org/10.1023/A:1005100221723.

———. 2002. "From "blame by praise" to "praise by blame": Analysis of verbal irony." *Journal of Pragmatics* 34 (6): 743–762.

Arnold, Magda B. 1960. *Emotion and Personality: Psychological Aspects.* New York: Columbia University Press.

Artstein, Ron, and Massimo Poesio. 2008. "Inter-Coder Agreement for Computational Linguistics." *Computational Linguistics* 34 (4): 555–596.

Atrey, Pradeep K., M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. 2010. "Multimodal Fusion for Multimedia Analysis: A Survey." *Multimedia Systems* 16 (6): 345–379. https://doi.org/10.1007/s00530-010-0182-0.

Attardo, Salvatore, Jodi Eisterhold, Jennifer Hay, and Isabella Poggi. 2003. "Multimodal markers of irony and sarcasm." *Humor* 16 (2): 243–260.

Aylett, Ruth, and Marc Cavazza. 2001. "Intelligent Virtual Environments-A state-of-the-art report" (January).

Aylett, Ruth, and Michael Luck. 2000. "Applying Artificial Intelligence to Virtual Reality: Intelligent Virtual Environments." *Applied Artificial Intelligence* 14 (1): 3–32. https://doi.org/10.1080/088395100117142.

Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations." In *Advances in Neural Information Processing Systems.*

Baños, Rosa M., Cristina Botella, Mariano Alcañiz, Víctor Liaño, and Beatriz Rey. 2006. "Changing Induced Moods Via Virtual Reality." In *Virtual Reality,* edited by Nathaniel I. Durlach and Anne S. Mavor, 71–84. Springer. https://doi.org/10.1007/11755494_3.

Barrett, Lisa Feldman, Ajay B. Satpute, and Eliza Bliss-Moreau. 2019. "Perception, Experience, and the Construction of Emotion." *Annual Review of Psychology* 70:1–26.

Bartoli, Luca, Giulio Corradi, and Franca Garzotto. 2013. "Supporting Children's Speech Therapy Through a Tablet-Based Interactive 3D Game." In *Proceedings of the 12th International Conference on Interaction Design and Children,* 43–52. ACM. https://doi.org/10.1145/2485760.2485767.

Baumgartner, Thomas, Martina Esslen, and Lutz Jäncke. 2006. "From Emotion Perception to Emotion Experience: Emotions Evoked by Pictures and Classical Music." *International Journal of Psychophysiology* 60 (1): 34–43.

Bethesda Game Studios. 2017. *The Elder Scrolls V: Skyrim VR.* Commercial VR Game Release; supports voice input through system-level speech recognition and community-implemented VR voice command mods.

Bhosale, Swapnil, Abhra Chaudhuri, Alex Lee Robert Williams, Divyank Tiwari, Anjan Dutta, Xiatian Zhu, Pushpak Bhattacharyya, and Diptesh Kanojia. 2023. *Sarcasm in Sight and Sound: Benchmarking and Expansion to Improve Multimodal Sarcasm Detection.* arXiv: 2310.01430 [cs.CL]. https://arxiv.org/abs/2310.01430.

Bird, Steven, Ewan Klein, and Edward Loper. 2009. "Natural Language Processing with Python." In *Proceedings of the ACL Interactive Poster and Demonstration Sessions,* 69–72.

Boersma, Paul, and David Weenink. 2001. "Praat, a system for doing phonetics by computer." *Glot International* 5 (9/10): 341–345.

Bohn, Jürgen, and Wolfgang Krüger. 1995. "Speaker-Independent Word Recognition for Virtual Environments Using Neural Networks." In *Proceedings of the International Conference on Artificial Neural Networks,* 899–904.

Bolt, Richard A. 1980. "Put-That-There: Voice and Gesture at the Graphics Interface." *Computer Graphics* 14 (3): 262–270.

Bolyanatz, Valeria, Jaime Soto-Barba, and Lidia Guerrero. 2024. "Prosodic irony in Chilean Spanish." *Language and Speech.*

Borod, Joan C. 1992. "Interhemispheric and Intrahemispheric Control of Emotion: A Focus on Unilateral Brain Damage." *Journal of Consulting and Clinical Psychology* 60 (3): 339–348.

———. 1993. *Emotional Processing after Brain Damage.* Oxford University Press.

Bradley, Margaret M., and Peter J. Lang. 1994. "Measuring emotion: The Self-Assessment Manikin and the semantic differential." *Journal of Behavior Therapy and Experimental Psychiatry* 25:49–59. https://doi.org/10.1016/0005-7916(94)90063-9.

Braun, Angelika, and Anita Schmiedel. 2018. "The Phonetics of Ambiguity: A Study on Verbal Irony." In *Cultures and Traditions of Wordplay and Wordplay Research,* edited by Esme Winter-Froemel and Verena Thaler, 529–562. Berlin: De Gruyter. https://doi.org/10.1515/9783110586374-026.

Braun, Bettina, and Anna Schmiedel. 2018. "The phonetics of irony." *Journal of Phonetics* 71:56–76.

British Aerospace. 1987. *The Virtual Cockpit Project.* Technical Report. British Aerospace Advanced Flight Deck Program.

Brown, Penelope, and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage.* Cambridge: Cambridge University Press.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. "Language Models are Few-Shot Learners." In *Advances in Neural Information Processing Systems.*

Bruce, Jake, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yannis M. Assael, David Silver, et al. 2024. "Genie: Generative Interactive Environments." *arXiv preprint arXiv:2402.15391.*

Bryant, Gregory A., and Jean E. Fox Tree. 2002. "Recognizing verbal irony in spontaneous speech." *Metaphor and Symbol* 17 (2): 99–117.

———. 2005. "Is there an ironic tone of voice?" *Language and Speech* 48 (3): 257–277.

Burkhardt, Felix, Astrid Paeschke, Miriam Rolfes, Walter F. Sendlmeier, and Benjamin Weiss. 2005. "A Database of German Emotional Speech." *Proceedings of Interspeech.*

Burkhardt, Felix, Maximilian Schmitt, and Björn Schuller. 2018. "Detecting irony in speech: Acoustic features and classification." In *Proceedings of Interspeech,* 776–780.

Busso, Carlos, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. "IEMOCAP: Interactive Emotional Dyadic Motion Capture Database." *Language Resources and Evaluation* 42 (4): 335–359.

Busso, Carlos, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. 2017. "MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception." *IEEE Transactions on Affective Computing* 8 (1): 67–80. https://doi.org/10.1109/TAFFC.2016.2515617.

Cambria, Erik, Björn Schuller, Yunqing Xia, and Catherine Havasi. 2013. "New Avenues in Opinion Mining and Sentiment Analysis." *IEEE Intelligent Systems* 28 (2): 15–21. https://doi.org/10.1109/MIS.2013.30.

Camp, Elisabeth. 2012. "Sarcasm, Pretense, and the Semantics/Pragmatics Distinction." *Noûs* 46 (4): 587–634. https://doi.org/10.1111/j.1468-0068.2010.00822.x.

Candy, Linda. 2006. *Practice-Based Research: A Guide.* Sydney: University of Technology Sydney.

Cannon, Walter B. 1927. "The James-Lange Theory of Emotions: A Critical Examination and an Alternative Theory." *The American Journal of Psychology* 39 (1/4): 106–124.

Cannon, Walter B., and Philip Bard. 1929. "Bodily Changes in Pain, Hunger, Fear and Rage: An Account of Recent Researches into the Function of Emotional Excitement." *The Journal of Nervous and Mental Disease* 69 (1): 108–110.

Carvalho, Isabel, Hugo Gonçalo Oliveira, and Catarina Silva. 2022. "Sentiment Analysis in Portuguese Dialogues." In *Proceedings of IberSPEECH 2022,* 176–180. Also available as a conference paper. Granada, Spain, 14-16 November.

———. 2023. "The Importance of Context for Sentiment Analysis in Dialogues." Early access, published 17 August 2023, *IEEE Access* 11:–. https://doi.org/10.1109/ACCESS.2023.3304633.

Castro, Santiago, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. "Towards Multimodal Sarcasm Detection (An _Obviously_ Perfect Paper)." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL).*

Castro, Santiago, Devamanyu Hazarika, Soujanya Poria, and Roger Zimmermann. 2019. "Towards multimodal sarcasm detection." In *Proceedings of the AAAI Conference on Artificial Intelligence,* 961–968.

Cavazza, Marc, Stéphane Donikian, Marc Christie, Ulrike Spierling, Nicolas Szilas, Peter Vorderer, Tilo Hartmann, et al. 2008. "The IRIS Network of Excellence: Integrating Research in Interactive Storytelling." In *Interactive Storytelling,* 5334:14–19. Lecture Notes in Computer Science. Springer. https://doi.org/10.1007/978-3-540-89454-4_3.

Chauhan, Dushyant, Devamanyu Hazarika, Soujanya Poria, and Roger Zimmermann. 2020. "Multimodal sarcasm detection using sentiment and emotion analysis." *Proceedings of the 28th ACM International Conference on Multimedia,* 362–370.

Cheang, Henry S., and Marc D. Pell. 2008. "The sound of sarcasm." *Speech Communication* 50 (5): 366–381.

Chicco, Davide, and Giuseppe Jurman. 2020. "The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation." *BMC Genomics* 21 (1).

Chion, Michel. 1994. *Audio-Vision: Sound on Screen.* Translated by Claudia Gorbman. Foreword by Walter Murch. Originally published as *L'audio-vision: son et image au cinéma* (Paris: Nathan, 1990). New York: Columbia University Press. ISBN: 0-231-07899-4.

Chirico, Alice, Francesco Ferrise, Lorenzo Cordella, and Andrea Gaggioli. 2017. "Designing Awe in Virtual Reality: An Experimental Study." *Frontiers in Psychology* 8:2351. https://doi.org/10.3389/fpsyg.2017.02351. https://www.frontiersin.org/articles/10.3389/fpsyg.2017.02351/full.

Chollet, François, et al. 2015. *Keras.* https://keras.io. Accessed for neural network implementation.

Chryssanthou, Yiorgos, George Caridakis, Katerina Kabassi, et al. 2009. "Emotion-based Interaction in CALLAS Augmented Reality Art Installations." In *Proceedings of the International Conference on Virtual Systems and Multimedia,* 223–230. IEEE.

Clark, Herbert H., and Richard J. Gerrig. 1984. "On the Pretense Theory of Irony." *Journal of Experimental Psychology: General* 113 (1): 121–126.

Cohen, Jacob. 1960. "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement* 20 (1): 37–46.

Colebrook, Claire. 2003. *Irony.* London: Routledge.

Conneau, Alexis, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. "XNLI: Evaluating Cross-Lingual Sentence Representations." *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*

Copet, Jade, Alexandre Défossez, Itai Gat, Qingqing Huang, Gabriel Synnaeve, Yossi Adi, et al. 2023. "AudioCraft: A Simple and Controllable Music Generation Framework." Framework paper/tech report accompanying MusicGen/AudioCraft releases, *arXiv preprint arXiv:2306.05284,* https://arxiv.org/abs/2306.05284.

Copet, Jade, Alexandre Défossez, Itai Gat, Huy Nguyen, Gabriel Synnaeve, and Yossi Adi. 2023. "Simple and Controllable Music Generation." *arXiv preprint arXiv:2306.05284,* https://arxiv.org/abs/2306.05284.

"EU project develops affective interface technologies for new media." 2007. Last update: 20 December 2007, *CORDIS News (European Commission).*

Cowie, Roddy, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G. Taylor. 2001. "Emotion Recognition in Human-Computer Interaction." *IEEE Signal Processing Magazine* 18 (1): 32–80. https://doi.org/10.1109/79.911197.

Cruz-Neira, Carolina, Daniel J. Sandin, and Thomas A. DeFanti. 1993. "Surround-Screen Projection-Based Virtual Reality: The Design and Implementation of the CAVE." In *Proceedings of SIGGRAPH,* 135–142. ACM.

Cutler, Anne. 1974. "On saying what you mean without meaning what you say." *Papers from the Tenth Regional Meeting of the Chicago Linguistic Society,* 117–127.

Darwin -Ekman edition, Charles. 1998. *The Expression of the Emotions in Man and Animals.* Edited by Paul Ekman. Oxford University Press.

Dash, Adyasha, and Kat R. Agres. 2024. "AI-Based Affective Music Generation Systems: A Review of Methods, and Challenges." *ACM Computing Surveys* 56 (11): 1–34. https://doi.org/10.1145/3672554.

Davidson, Richard J. 1992. "Anterior Cerebral Asymmetry and the Nature of Emotion." *Brain and Cognition* 20 (1): 125–151.

———. 1993. "The Neuropsychology of Emotion and Affective Style." *Psychological Science* 4 (2): 71–76.

Davis, H., and S. Mohammad. 2014. "Generating Music from Literature." In *Proceedings of the AAAI Conference on Artificial Intelligence.*

Davitz, Joel R. 1964. *The Communication of Emotional Meaning.* McGraw-Hill.

Debevec, Paul E., Camillo J. Taylor, and Jitendra Malik. 1996. "Modeling and Rendering Architecture from Photographs: A Hybrid Geometry- and Image-Based Approach." In *Proceedings of SIGGRAPH 1996,* 11–20. ACM. https://doi.org/10.1145/237170.237191.

Deleuze, Gilles, and Félix Guattari. 1983. *Anti-Oedipus: Capitalism and Schizophrenia.* Translated by Robert Hurley, Mark Seem, and Helen R. Lane. Original work published 1972. University of Minnesota Press.

———. 1987. *A Thousand Plateaus: Capitalism and Schizophrenia.* Translated by Brian Massumi. University of Minnesota Press.

Demszky, Dorottya, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. "GoEmotions: A Dataset of Fine-Grained Emotions." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020),* 4040–4054. Online.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of NAACL-HLT,* 4171–4186. https://doi.org/10.18653/v1/N19-1423.

Dhariwal, Prafulla, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. 2020. "Jukebox: A Generative Model for Music." *arXiv preprint arXiv:2005.00341,* https://arxiv.org/abs/2005.00341.

Dionisio, John David N., William G. Burns III, and Richard Gilbert. 2013. "3D Virtual Worlds and the Metaverse: Current Status and Future Possibilities." *ACM Computing Surveys* 45 (3): 34:1–34:38. https://doi.org/10.1145/2480741.2480751.

Dozio, Nicolò, Federica Marcolin, Giulia Wally Scurati, Francesco Ferrise, Giuseppe Riva, and Mariano Alcañiz. 2021. "Development of an Affective Database Made of Interactive Virtual Environments." *Scientific Reports* 11:23257. https://doi.org/10.1038/s41598-021-02781-6.

Dozio, Nicolò, Federica Marcolin, Giulia Wally Scurati, and Francesco Ferrise. 2022. "A Design Methodology for Affective Virtual Reality Based on the Valence–Arousal–Dominance Model." *International Journal of Human–Computer Studies* 165:102918. https://doi.org/10.1016/j.ijhcs.2022.102918.

Duffy, Elizabeth. 1957. "The Psychological Significance of the Concept of "Arousal" or "Activation"." *Psychological Review* 64 (5): 265–275. https://doi.org/10.1037/h0048837. https://doi.org/10.1037/h0048837.

Egger, M., M. Ley, and S. Hanke. 2019. "Emotion recognition from physiological signal analysis: A review." *Electronic Notes in Theoretical Computer Science* 343:35–55.

Eisenstein, Sergei M., Vsevolod I. Pudovkin, and Grigori V. Alexandrov. 1985. "A Statement on Sound." In *Film Sound: Theory and Practice,* edited by Elisabeth Weis and John Belton, 83–85. Originally published 1928. New York: Columbia University Press.

Ekman, Paul. 1992. "An Argument for Basic Emotions." *Cognition and Emotion* 6 (3–4): 169–200.

Ekman, Paul. 2005. "Basic Emotions." In *Handbook of Cognition and Emotion,* edited by Tim Dalgleish and Mick Power, 45–60. John Wiley & Sons.

Ekman, Paul, and Wallace V. Friesen. 1971. "Constants across cultures in the face and emotion." *Journal of Personality and Social Psychology* 17 (2): 124–129.

El Ayadi, Moataz, Mohamed S. Kamel, and Fakhri Karray. 2011. "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases." *Pattern Recognition* 44 (3): 572–587.

Elizalde, Benjamin, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2022. "CLAP: Learning Audio Concepts From Natural Language Supervision." *arXiv preprint arXiv:2206.04769,* https://arxiv.org/abs/2206.04769.

Esuli, Andrea, and Fabrizio Sebastiani. 2006. "SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining." *Proceedings of the International Conference on Language Resources and Evaluation.*

Evans, Zach, CJ Carr, Josiah Taylor, Scott H. Hawley, and Jordi Pons. 2024. "Fast Timing-Conditioned Latent Audio Diffusion." *arXiv preprint arXiv:2402.04825,* https://arxiv.org/abs/2402.04825.

Eyben, Florian, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Devillers, et al. 2016. "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing." *IEEE Transactions on Affective Computing* 7 (2): 190–202.

Eyben, Florian, Björn Schuller, Martin Wöllmer, Robert Friesen, and Gerhard Rigoll. 2013. "The Geneva minimalistic acoustic parameter set." In *Proceedings of the 3rd International Workshop on Audio/Visual Emotion Challenge,* 1–6.

Eyben, Florian, Martin Wöllmer, and Björn Schuller. 2010. "openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor." In *Proceedings of the 18th ACM International Conference on Multimedia (MM'10),* 1459–1462. Firenze, Italy: ACM. https://doi.org/10.1145/1873951.1874246.

Eysenck, Michael W., and Mark T. Keane. 2000. *Cognitive Psychology: A Student's Handbook.* 4th. Psychology Press.

Felnhofer, Anna, Oswald D. Kothgassner, Mareike Schmidt, Anna-Katharina Heinzle, Leon Beutl, Helmut Hlavacs, and Ivo Kryspin-Exner. 2015. "Is Virtual Reality Emotionally Arousing? Investigating Five Emotion-Inducing Virtual Park Scenarios." *International Journal of Human–Computer Studies* 82:48–56. https://doi.org/10.1016/j.ijhcs.2015.05.004.

Ferreira, Lucas N., Levi H. S. Lelis, and Jim Whitehead. 2020. "Computer-Generated Music for Tabletop Role-Playing Games." In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE),* 16:59–65. 1. https://doi.org/10.1609/aiide.v16i1.7408.

Ferreira, Lucas N., and Jim Whitehead. 2019. "Learning to Generate Music with Sentiment." In *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR),* 384–390.

Fisher, Scott S., Mike McGreevy, J. Humphries, and Warren Robinett. 1989. *The Virtual Interface Environment Workstation (VIEW).* Technical report NASA-TM-102203. NASA Ames Research Center.

Fleiss, Joseph L. 1971. "Measuring Nominal Scale Agreement among Many Raters." *Psychological Bulletin* 76 (5): 378–382. https://doi.org/10.1037/h0031619.

Fónagy, Iván. 1971. *La vive voix.* Paris: Payot.

Forero, Jorge, Gilberto Bernardes, and Mónica Mendes. 2023. *Are Words Enough? On the semantic conditioning of affective music generation.* arXiv: 2311.03624 [cs.MM]. https://arxiv.org/abs/2311.03624.

Forsgren, Seth, and Hayk Martiros. 2022. *Riffusion: Stable Diffusion for Real-Time Music Generation.* https://www.riffusion.com/about. Accessed 2026-01-02.

Fredrickson, Barbara L. 1998. "What Good Are Positive Emotions?" *Review of General Psychology* 2 (3): 300–319. https://doi.org/10.1037/1089-2680.2.3.300.

Frijda, Nico H. 1986. *The Emotions.* Cambridge: Cambridge University Press.

Frijda, Nico H., Antony S. R. Manstead, and Keith Oatley. 1989. *The Emotions.* Cambridge University Press.

Furness, Thomas A. 1986. "The Super Cockpit and Its Human Factors Challenges." *Proceedings of the IEEE* 74 (4): 528–538.

Gabrielsson, A., and E. Lindström. 2001. "The Influence of Musical Structure on Emotional Expression." *Music and Emotion: Theory and Research.*

García-Hernández, Rosa A., et al. 2024. "A Systematic Literature Review of Modalities, Trends, and Limitations in Emotion Recognition, Affective Computing, and Sentiment Analysis." *Applied Sciences* 14 (16): 7165. https://doi.org/10.3390/app14167165.

Gibbs, Raymond W. 2000. "Irony in Talk Among Friends." *Metaphor and Symbol* 15 (1–2): 5–27. https://doi.org/10.1080/10926488.2000.9678862.

Gibbs, Raymond W., and Herbert L. Colston. 2012. *Interpreting Figurative Meaning.* Cambridge: Cambridge University Press.

Gilroy, Stephen W., Hartmut Seichter, Maurice Benayoun, Marc Cavazza, Rémi Chaignon, Satu-Marja Mäkelä, Markus Niiranen, et al. 2008. "E-Tree: Emotionally Driven Augmented Reality Art." In *Proceedings of the 16th ACM International Conference on Multimedia (MM '08),* 945–948. ACM. https://doi.org/10.1145/1459359.1459529.

González-Fuente, Santiago, Pilar Prieto, and Verónica López. 2016. "The prosody of irony in French." *Journal of Phonetics* 59:1–18.

Gorbman, Claudia. 1987. *Unheard Melodies: Narrative Film Music.* Bloomington and London: Indiana University Press / BFI Publishing. ISBN: 0-253-20436-4.

Gosselin, Pierre, Gilles Kirouac, and Françoise Y. Doré. 2005. "Components and Recognition of Facial Expressions in the Communication of Emotion by Actors." *Journal of Personality and Social Psychology* 89 (5): 705–718.

Granato, Marco, Daniela Gadia, and Gustavo Marfia. 2020. "Emotion Recognition in Virtual Reality Racing Games Through Physiological Signal Analysis." *Sensors* 20 (18): 5163. https://doi.org/10.3390/s20185163.

Grandjean, Didier, David Sander, and Klaus R. Scherer. 2008. "Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization." *Consciousness and Cognition* 17 (2): 484–495. https://doi.org/10.1016/j.concog.2008.03.019.

Grice, H. Paul. 1975. "Logic and Conversation." In *Syntax and Semantics, Vol. 3: Speech Acts,* edited by Peter Cole and Jerry L. Morgan, 41–58. New York: Academic Press.

Gu, Simeng, Fei Wang, Nitesh P. Patel, James A. Bourgeois, and Jason H. Huang. 2019. "A Model for Basic Emotions Using Observations of Behavior in Drosophila." *Frontiers in Psychology* 10:781. https://doi.org/10.3389/fpsyg.2019.00781.

Gu, Simeng, Fei Wang, Ting Yuan, Baohua Guo, and Jason H. Huang. 2015. "Differentiation of Primary Emotions through Neuromodulators: Review of Literature." *International Journal of Neurology* 1:43–50.

Guo, Yuwei, Ceyuan Yang, Anyi Rao, Zhengyang Geng, Yixiao Ge, Yining Li, Jianbo Shi, and Ying Shan. 2023. "AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning." *arXiv preprint arXiv:2307.04725.*

Halpern, Ariel, Lorraine Ramig, Shimon Sapir, Thomas F. Quatieri, and Nathan Ward. 2019. "Innovative Speech and Language Therapy Through Virtual Reality." *Journal of Speech, Language, and Hearing Research* 62 (9): 3381–3392. https://doi.org/10.1044/2019_JSLHR-S-18-0444.

Haverkate, Henk. 1990. *A Speech Act Analysis of Irony.* Amsterdam: John Benjamins.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. "Deep Residual Learning for Image Recognition." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 770–778. https://doi.org/10.1109/CVPR.2016.90.

Ho, Jonathan, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. 2022. "Imagen Video: High Definition Video Generation with Diffusion Models." *arXiv preprint arXiv:2210.02303.*

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation* 9 (8).

Hofmann, S. M., F. Klotzsche, A. Mariola, V. Nikulin, A. Villringer, and M. Gaebler. 2021. "Decoding subjective emotional arousal from EEG during an immersive virtual reality experience." *eLife* 10:e64812.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. "spaCy: Industrial-strength Natural Language Processing in Python." *Zenodo,* https://doi.org/10.5281/zenodo.1212303.

Hsu, Wei-Ning, Benjamin Bolte, Yao-Huang Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units." In *Proceedings of the Annual Meeting of the Association for Computational Linguistics.*

Hu, Minqing, and Bing Liu. 2004. "Mining and Summarizing Customer Reviews." *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Huang, Chih-Fang, and Cheng-Yuan Huang. 2020. "Emotion-based AI Music Generation System with CVAE-GAN." In *2020 IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE),* 220–222. IEEE. https://doi.org/10.1109/ECICE50847.2020.9301934.

Huang, Qingqing, Aren Jansen, Joonseok Lee, Raghav Ganti, Jitong Li, and Daniel P. W. Ellis. 2022. "Generating Melody with Emotion Constraint." In *Proceedings of the ACM International Conference on Multimedia Retrieval (ICMR).* ACM. https://doi.org/10.1145/3501409.3501691.

Huang, Qingqing, Daniel S. Park, Tao Wang, Timo I. Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, et al. 2023. "Noise2Music: Text-Conditioned Music Generation with Diffusion Models." *arXiv preprint arXiv:2302.03917,* https://arxiv.org/abs/2302.03917.

Huang, Ji-Shi. 2013. *T(w)peeper.* Interactive generative media artwork. Sentiment-driven visualization of live Twitter data. https://jiesihuang.com/twpeeper/.

Huskisson, E. C. 1974. "Measurement of Pain." *The Lancet* 304 (7889): 1127–1131.

Ireland, David. 2017. "Great Expectations? The Changing Role of Audiovisual Incongruence in Contemporary Multimedia." *Music and the Moving Image* 10 (3): 21–35. ISSN: 2167-8464. https://doi.org/10.5406/musimoviimag.10.3.0021.

———. 2018. *Identifying and Interpreting Incongruent Film Music.* Palgrave Studies in Audio-Visual Culture. Cham: Palgrave Macmillan. ISBN: 978-3-030-00505-4. https://doi.org/10.1007/978-3-030-00506-1.

Issa, Darin, Mustafa Fatih Demirci, and Ali Yazici. 2020. "Speech Emotion Recognition Using Deep Convolutional Neural Networks and Mel-Frequency Cepstral Coefficients." *Multimedia Tools and Applications* 79 (21): 15599–15627.

Izard, Carroll E. 1977. *Human Emotions.* Plenum Press.

———. 1991. *The Psychology of Emotions.* Plenum Press.

———. 2007. "Basic Emotions, Natural Kinds, Emotion Schemas, and a New Paradigm." *Perspectives on Psychological Science* 2 (3): 260–280.

Jack, Rachael E., Oliver G. B. Garrod, and Philippe G. Schyns. 2014. "Dynamic Facial Expressions of Emotion Transmit an Evolving Hierarchy of Signals over Time." *Current Biology* 24 (2): 187–192. https://doi.org/10.1016/j.cub.2013.11.064.

Jaiswal, Mimansa, Zhi Ding, Wen Chen, Soujanya Poria, and Louis-Philippe Morency. 2020. "MuSE: a Multimodal Dataset of Stressed Emotion." In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC).*

James, William. 1884. "What is an Emotion?" *Mind* 9 (34): 188–205.

Joachims, Thorsten. 1997. "A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization." In *Proceedings of the 14th International Conference on Machine Learning.* Morgan Kaufmann. https://doi.org/10.5555/645526.657278.

Johnson, W. Lewis, Stacy Marsella, and Hannes Vilhjálmsson. 2004. "The Intelligent Virtual Environment for Language Learning (IVELL)." In *Proceedings of the International Conference on Intelligent Virtual Agents,* 219–221. Lecture Notes in Computer Science. Springer.

Johnson, W. Lewis, Jeff Rickel, and James C. Lester. 1998. "Animated Pedagogical Agents: Face-to-Face Interaction in Interactive Learning Environments." In *International Journal of Artificial Intelligence in Education,* 11:47–78. 47.

Jolliffe, Ian T., and Jorge Cadima. 2016. *Principal Component Analysis.* 2nd ed. Springer.

Jorgensen, Julia. 1996. "The Functions of Sarcastic Irony in Speech." *Journal of Pragmatics* 26 (5): 613–634. https://doi.org/10.1016/0378-2166(95)00067-4.

Junior, Alfredo Pereira, and Fei Wang. 2016. "Neuromodulation, Emotional Feelings and Affective Disorders." *Mens Sana Monographs* 14 (1): 53–70. https://doi.org/10.4103/0973-1229.193073.

Jurafsky, Dan, and James H. Martin. 2019. *Speech and Language Processing.* 3rd ed. Pearson.

Juslin, Patrik N. 2013. "From everyday emotions to aesthetic emotions: Towards a unified theory of musical emotions." *Physics of Life Reviews.*

Juslin, Patrik N., and John A. Sloboda, eds. 2010. *Handbook of Music and Emotion: Theory, Research, Applications.* Oxford University Press.

Kerlinger, Fred N. 1973. *Foundations of Behavioral Research.* Holt, Rinehart / Winston.

Keung, Phillip, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. "Multilingual Amazon Reviews Corpus." *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*

Khalaf, OI, D Srinivasan, S Algburi, J Vellaichamy, D Selvaraj, MS Sharif, and W Elmedany. 2024. "Elevating metaverse virtual reality experiences through network-integrated neuro-fuzzy emotion recognition and adaptive content generation algorithms." *Engineering Reports* 6 (e12894). https://doi.org/10.1002/eng2.12894.

Khodak, Mikhail, Nikunj Saunshi, and Kiran Vodrahalli. 2018. "A Large Self-Annotated Corpus for Sarcasm." *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*

Kim, SunJung, and Elisabeth André. 2004. "Composing Affective Music with a Generate and Sense Approach." In *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS),* 316–321.

Kim, Yoon. 2014. "Convolutional Neural Networks for Sentence Classification." *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*

Kipp, Michael. 2001. "ANVIL - a generic annotation tool for multimodal dialogue." In *Interspeech.* https://api.semanticscholar.org/CorpusID:15124598.

Kittler, Josef, Mohamad Hatef, Robert P. W. Duin, and Jiří Matas. 1998. "Combining Classifiers: A Theoretical Framework." *Pattern Analysis and Applications* 1 (1): 18–27. https://doi.org/10.1007/BF01238023.

Kleinen, T. 1968. "Semantic Differential Studies of Emotion." *Psychological Research* 31:24–36.

Knox, Norman D. 1961. *The Word Irony and Its Context: 1500–1755.* Durham, NC: Duke University Press.

Koelstra, Sander, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2012. "DEAP: A database for emotion analysis using physiological signals." *IEEE Transactions on Affective Computing* 3 (1): 18–31.

Kossaifi, Jean, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, et al. 2019. "SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild." *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

Kreuz, Roger J., and Sam Glucksberg. 1998. "How to Be Sarcastic: The Echoic Reminder Theory of Verbal Irony." *Journal of Experimental Psychology: General* 127 (4): 374–386. https://doi.org/10.1037/0096-3445.127.4.374.

Kreuz, Roger J., and Richard M. Roberts. 1995. "Two cues for verbal irony: Hyperbole and the ironic tone of voice." *Metaphor and Symbol* 10 (1): 21–31.

Krippendorff, Klaus. 2004. "Reliability in Content Analysis: Some Common Misconceptions and Recommendations." *Human Communication Research* 30 (3): 411–433. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x.

Krugmann, Jan Ole, and Jochen Hartmann. 2024. "Sentiment Analysis in the Age of Generative AI." *Customer Needs and Solutions* 11 (1): 1–19. https://doi.org/10.1007/s40547-024-00143-4. https://doi.org/10.1007/s40547-024-00143-4.

Kumar, Shivani, Shubham Garg, and Amit Singh. 2024. "SemEval-2024 Task 10: Emotion Discovery and Reasoning its Flip in Conversation (EDiReF)." ArXiv:2402.04718, *arXiv preprint,* https://arxiv.org/abs/2402.04718.

Laird, John E., Paul S. Rosenbloom, and Allen Newell. 1987. "Soar: An Architecture for General Intelligence." *Artificial Intelligence* 33 (1): 1–64.

Lang, Peter J. 1993. "The Network Model of Emotion: Motivational Connections." *Nebraska Symposium on Motivation* 41:1–38.

Lange, Carl Georg. 1885. *The Emotions.* English translation. Original work published in Danish (1885). Williams & Norgate. https://ia801604.us.archive.org/4/items/emotions00lang/emotions00lang.pdf.

Lazarus, Richard S. 1991. "Emotion and Adaptation." In *The Handbook of Personality: Theory and Research,* edited by Lawrence A. Pervin, 609–637. New York: Guilford Press.

Lee, Chul Min, and Shrikanth S. Narayanan. 2005. "Toward Detecting Emotions in Spoken Dialogs." *IEEE Transactions on Speech and Audio Processing* 13 (2): 293–303. https://doi.org/10.1109/TSA.2004.838534.

Lee, Lik-Hang, Tristan Braud, Pengyuan Zhou, Lin Wang, Dianlei Xu, Zijun Lin, Abhishek Kumar, Carlos Bermejo, and Pan Hui. 2021. "All One Needs to Know about Metaverse: A Complete Survey on Technological Singularity, Virtual Ecosystem, and Research Agenda." *Journal of Latex Class Files,* arXiv: 2110.05352.

Levin, Golan, and Zachary Lieberman. 2004. “The Hidden Worlds of Noise and Voice.” In *Ars Electronica Festival: Timeshift.* Interactive AR installation exploring real-time voice visualization. Ars Electronica Center.

Li, Jianfeng, Yichi Zhang, Rui Wang, and Liang Chen. 2021. “A 360-Degree Immersive Video Database for Emotional Analysis Based on Valence-Arousal Model.” *IEEE Access* 9:123456–123468. https://doi.org/10.1109/ACCESS.2021.XXXXXXX.

Li, Shu, and Weiping Deng. 2020. “Deep learning for emotion recognition from facial expressions: A survey.” *IEEE Transactions on Affective Computing.*

Li, Yifan, Soujanya Poria, and Devamanyu Hazarika. 2024. “Prosody–semantics interaction in sarcastic speech.” *Speech Communication.*

Lin, Chen-Hsuan, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros, et al. 2023. “Magic3D: High-Resolution Text-to-3D Content Creation.” In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Liu, Bing. 2012. *Sentiment Analysis and Opinion Mining.* Morgan & Claypool.

———. 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions.* Cambridge University Press.

Liu, Ruoshi, Tinghui Zhou, Alexei A. Efros, et al. 2023. “Zero-1-to-3: Zero-Shot One Image to 3D Object.” In *IEEE International Conference on Computer Vision (ICCV).*

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. “RoBERTa: A Robustly Optimized BERT Pretraining Approach.” *arXiv preprint arXiv:1907.11692.*

Livingstone, Steven, Ralf Muhlberger, Andrew Brown, and William Thompson. 2010. “Changing Musical Emotion: A Computational Rule System for Modifying Score and Performance.” *Computer Music Journal* 34 (March): 41–64. https://doi.org/10.1162/comj.2010.34.1.41.

Livingstone, Steven R., and Frank A. Russo. 2018. “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS).” *PLOS ONE* 13 (5).

Lotfian, Reza, and Carlos Busso. 2019. “Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech from Existing Podcast Recordings.” *IEEE Transactions on Affective Computing* 10 (4): 471–483.

Lucas, Gale M., Jonathan Gratch, Andrew King, and Louis-Philippe Morency. 2014. “It’s Only a Computer: Virtual Humans Increase Willingness to Disclose.” *Computers in Human Behavior* 37:94–100. https://doi.org/10.1016/j.chb.2014.04.043.

Ma, Ziyang, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2023. “emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation.” *arXiv preprint arXiv:2312.15185.*

Maas, Andrew L., Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. “Learning Word Vectors for Sentiment Analysis.” *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL),* 142–150. https://aclanthology.org/P11-1015/.

Madhok, Rishi, Shubham Goel, and Shreya Garg. 2018. "SentiMozart: Music Generation based on Emotions." In *Proceedings of the 10th International Conference on Agents and Artificial Intelligence (ICAART),* 2:501–506. SCITEPRESS.

Malpica, Natalia, et al. 2023. "Measuring the Influence of Audio on Immersive Experience in Extended Reality and Digital Games: A Systematic Review." Study on audiovisual thematic cohesion in VR/games environments, examining the effect of audiovisual thematic dissonance on player experience and presence. Available at: https://www.researchgate.net/publication/374938679, *ResearchGate Preprint / Conference Proceedings.*

Mandler, George. 1975. *Mind and Emotion.* New York: John Wiley & Sons.

———. 1984. *Mind and Body: Psychology of Emotion and Stress.* New York: W. W. Norton.

Maples-Keller, Jessica L., Brian E. Bunnell, Seokmin J. Kim, and Barbara O. Rothbaum. 2017. "The Use of Virtual Reality Technology in the Treatment of Anxiety and Other Psychiatric Disorders." *Harvard Review of Psychiatry* 25 (3): 103–113. https://doi.org/10.1097/HRP.0000000000000138.

Marín-Morales, Jorge, Fabio Pallavicini, Mateu Sbert, Giuseppe Riva, and Mariano Alcañiz. 2018. "Emotional response to immersive virtual environments: A comparison between virtual reality and 360° videos." *Frontiers in Psychology* 9:1500. https://doi.org/10.3389/fpsyg.2018.01500.

Martinho, Carlos, and Ana Paiva. 2000. "Using Anticipation to Create Believable Behaviour." In *Proceedings of the International Conference on Autonomous Agents,* 148–155. Emotion-driven virtual characters with appraisal-based narrative influence. ACM. https://doi.org/10.1145/336595.336617.

Mauchand, Maël, Jonathan A. Caballero, Xiaoming Jiang, and Marc D. Pell. 2021. "Immediate Online Use of Prosody Reveals the Ironic Intentions of a Speaker: Neurophysiological Evidence." *Cognitive, Affective, & Behavioral Neuroscience* 21 (1): 74–92. https://doi.org/10.3758/s13415-020-00849-7.

McArthur, Victoria. 2016. "Mismatched Audio-Visual Cues and Immersion in Cinematic Virtual Reality." As cited in: Jones, S. et al. (2023). Mapping the viewer experience in cinematic virtual reality: A systematic review. *Presence: Virtual and Augmented Reality*. https://doi.org/10.1162/pres_a_00409, *Presence: Teleoperators and Virtual Environments.*

McCormack, Heather M., David J. Horne, and Simon Sheather. 1988. "Clinical Applications of Visual Analogue Scales: A Critical Review." *Psychological Medicine* 18 (4): 1007–1019.

McGlashan, Scott. 1995. "A Spoken Language Controlled Virtual Environment." In *Proceedings of the ESCA Workshop on Spoken Dialogue Systems,* 165–168.

———. 1996. "Using Speech to Control Agents in a Virtual Environment." *Speech Communication* 18 (3): 255–267.

McKeown, Gary, Michel F. Valstar, Roddy Cowie, Maja Pantic, and Marc Schröder. 2012. "The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent." *IEEE Transactions on Affective Computing* 3 (1): 5–17. https://doi.org/10.1109/T-AFFC.2011.20.

McNair, Douglas M., Maurice Lorr, and Leo F. Droppleman. 1971. *Manual for the Profile of Mood States.* Educational / Industrial Testing Service.

Mehrabian, Albert, and Susan R. Ferris. 1967. “Inference of Attitudes from Nonverbal Communication in Two Channels.” *Journal of Consulting Psychology* 31 (3): 248–252. https://doi.org/10.1037/h0024648.

Mehrabian, Albert, and Morton Wiener. 1967. “Decoding of Inconsistent Communications.” *Journal of Personality and Social Psychology* 6 (1): 109–114. https://doi.org/10.1037/h0024532.

Mescheder, Lars, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. 2019. “Occupancy Networks: Learning 3D Reconstruction in Function Space.” In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Meuleman, Bastien, David Rudrauf, and Jean-Baptiste Lamy. 2018. “A Virtual Reality Framework for Emotion Elicitation and Recognition Based on Interactive Gameplay.” In *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces,* 1–8. IEEE.

Mildenhall, Ben, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. “NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis.” In *European Conference on Computer Vision (ECCV).* Springer.

Mirsamadi, Seyedmahdad, Emad Barsoum, and Cha Zhang. 2017. “Automatic Speech Emotion Recognition Using Recurrent Neural Networks with Local Attention.” In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

Mohammad, Saif M., Svetlana Kiritchenko, and Xiaodan Zhu. 2013. “NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets.” *Proceedings of the International Workshop on Semantic Evaluation.*

Mordvintsev, Alexander, Christopher Olah, and Mike Tyka. 2015. *Inceptionism: Going Deeper into Neural Networks.* https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html. Google Research Blog post introducing DeepDream.

Mott, Bradford, Sungchul Lee, and James Lester. 2019. “Conversational AI for Interactive Games.” *AI Magazine* 40 (1): 46–55.

Mousavi, Elahe, Alfredo Vellido, and Antonios Liapis. 2024. “Magic XRoom: Flow-Driven Affective Modulation in Adaptive Virtual Environments.” *IEEE Transactions on Affective Computing,* https://doi.org/10.1109/TAFFC.2024.XXXXXX.

Muecke, D. C. 1969. *The Compass of Irony.* Methuen.

Müller, Thomas, Alex Evans, Christoph Schied, and Alexander Keller. 2022. “Instant Neural Graphics Primitives with a Multiresolution Hash Encoding.” *ACM Transactions on Graphics* 41 (4): 102:1–102:15. https://doi.org/10.1145/3528223.3530127.

Nesse, Randolph M. 1990. “Evolutionary Explanations of Emotions.” *Human Nature* 1 (3): 261–289. https://doi.org/10.1007/BF02733986.

Niebuhr, Oliver. 2014. “On the phonetics of sarcasm.” *Proceedings of Speech Prosody.*

Nielzén, S., and Z. Cesarec. 1981. “Emotional Meaning of Music as Measured by the Semantic Differential.” *Psychology of Music* 9 (2): 47–56.

Nomura, Kazunori, and Michitaka Fukumoto. 2018. “Music Melodies Suited to Multiple Users’ Feelings Composed by Asynchronous Distributed Interactive Genetic Algorithm.” *International Journal of Software Innovation* 6 (1): 26–36.

Ochs, Magalie, and Catherine Pelachaud. 2017. "An Intelligent Virtual Patient for Training Doctors in Breaking Bad News." *Proceedings of the International Conference on Intelligent Virtual Agents,* 38–47.

Osgood, Charles E. 1953. *Method and Theory in Experimental Psychology.* Oxford University Press.

Osgood, Charles E., George J. Suci, and Percy H. Tannenbaum. 1957. *The Measurement of Meaning.* University of Illinois Press.

Oviatt, Sharon. 1999. "Ten Myths of Multimodal Interaction." *Communications of the ACM* 42 (11): 74–81.

Panda, Renato, Ricardo Malheiro, and Rui Pedro Paiva. 2018. "Novel Audio Features for Music Emotion Recognition." *IEEE Transactions on Affective Computing* 11 (4): 637–647. https://doi.org/10.1109/TAFFC.2018.2820691.

Pang, Bo, and Lillian Lee. 2004. "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts." *Proceedings of the Annual Meeting of the Association for Computational Linguistics.*

———. 2008. "Opinion Mining and Sentiment Analysis." *Foundations and Trends in Information Retrieval* 2 (1–2).

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. "Thumbs Up? Sentiment Classification Using Machine Learning Techniques." *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*

Panksepp, Jaak. 1982. "Toward a General Psychobiological Theory of Emotions." *Behavioral and Brain Sciences* 5 (3): 407–422.

———. 1998. *Affective Neuroscience: The Foundations of Human and Animal Emotions.* Oxford University Press.

Park, Jeong Joon, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. 2019. "DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation." In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Parkinson, Brian. 1995. *Ideas and Realities of Emotion.* Routledge.

Parrott, W. Gerrod. 2001. *Emotions in Social Psychology: Essential Readings.* Psychology Press.

Peng, Kuan-Chuan, Tsuhan Chen, Amir Sadovnik, and Andrew Gallagher. 2015. "A mixed bag of emotions: Model, predict, and transfer emotion distributions," 860–868. June. https://doi.org/10.1109/CVPR.2015.7298687.

Pennebaker, James W., Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The Development and Psychometric Properties of LIWC2015.* University of Texas at Austin.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. "GloVe: Global Vectors for Word Representation." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),* 1532–1543. Doha, Qatar: Association for Computational Linguistics, October. https://doi.org/10.3115/v1/D14-1162. https://aclanthology.org/D14-1162.

Pepino, Leonardo, Pablo Riera, and Luciana Ferrer. 2021. "Emotion Recognition from Speech Using wav2vec 2.0 Embeddings." *arXiv preprint arXiv:2104.03502.*

Peretz, Isabelle, Louise Gagnon, and Bernard Bouchard. 1997. "Music and Emotion: Perceptual Determinants, Immediacy, and Isolation after Brain Damage." *Cognition* 68 (2): 111–141.

Picard, Rosalind W. 1997. *Affective Computing.* MIT Press.

Pinilla, Andres, Jaime Garcia, William Raffe, Jan-Niklas Voigt-Antons, Robert P Spang, and Sebastian Möller. 2021. "Affective visualization in Virtual Reality: An integrative review." *Frontiers in Virtual Reality,* 1–27.

Plutchik, Robert. 1960. "The Multifactor-Analytic Theory of Emotion." *Journal of Psychology* 50:153–171.

———. 1980. *Emotion: A Psychoevolutionary Synthesis.* Harper & Row.

———. 1982. *Emotion: A Psychoevolutionary Synthesis.* Harper & Row.

Pontiki, Maria, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2016. "SemEval-2016 Task 5: Aspect-Based Sentiment Analysis." *Proceedings of the International Workshop on Semantic Evaluation.*

Poole, Ben, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. 2023. "DreamFusion: Text-to-3D Using 2D Diffusion." In *International Conference on Learning Representations (ICLR).*

Poria, Soujanya, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. "MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations." *Proceedings of the Annual Meeting of the Association for Computational Linguistics.*

Posner, Jonathan, James A. Russell, and Bradley S. Peterson. 2005. "The Circumplex Model of Affect: An Integrative Approach to Affective Neuroscience, Cognitive Development, and Psychopathology." *Development and Psychopathology* 17 (3): 715–734.

Potts, Christopher, Zheng Ye, Zheng Hu, Yada Liu, and Yiyang Zhang. 2021. "DynaSent: A Dynamic Benchmark for Sentiment Analysis." *Proceedings of the Annual Meeting of the Association for Computational Linguistics.*

Price, Donald D., Patricia A. McGrath, Abbas Rafii, and Brian Buckingham. 1983. "The Validation of Visual Analogue Scales as Ratio Scale Measures for Chronic and Experimental Pain." *Pain* 17 (1): 45–56.

Pumarola, Albert, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. 2021. "D-NeRF: Neural Radiance Fields for Dynamic Scenes." In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Quintilian. 1921. *The Institutio Oratoria.* Edited by H. E. Butler. Cambridge, MA: Harvard University Press.

Radford, Alec, Jong Wook Kim, et al. 2023. "Robust Speech Recognition via Large-Scale Weak Supervision." *arXiv:2212.04356.*

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. "Robust Speech Recognition via Large-Scale Weak Supervision." In *Proceedings of the International Conference on Machine Learning.*

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. *Improving Language Understanding by Generative Pre-Training.* Technical report. OpenAI.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. *Language Models are Unsupervised Multitask Learners.* Technical report. OpenAI.

Rakov, Rafal, and Andrew Rosenberg. 2013. "Sure, I did the right thing: A system for sarcasm detection in speech." *Proceedings of Interspeech,* 842–846.

Rao, Rajiv, Rebecca Ronquest, and Lauren Hall-Lew. 2022. "Prosodic realization of sarcasm in bilingual speech." *Journal of Sociolinguistics.*

Ray, Parthasarathi, Pradyumna Mishra, and Soujanya Poria. 2022. "MUStARD++: Multimodal sarcasm detection with expanded annotations." *IEEE Transactions on Affective Computing.*

Reddy, D. Raj. 1976. "Speech Recognition by Machine: A Review." *Proceedings of the IEEE* 64 (4): 501–531.

Reilly, W. Scott. 1996. "Believable Social and Emotional Agents." In *Technical Report CMU-CS-96-138.* Oz Project; appraisal-based emotional agents grounded in the OCC model. Pittsburgh, PA.

Ringeval, Fabien, Andreas Sonderegger, Jürgen Sauer, and Denis Lalanne. 2013. "Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions." In *Proceedings of the 2nd International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE).*

Rizzo, Albert, Jonathan Gratch, Stephen Woods, and et al. 2014. "SimCoach: An Online Virtual Human Support System for Service Members and Veterans." In *Proceedings of the 14th International Conference on Intelligent Virtual Agents,* 424–427. Lecture Notes in Computer Science. Springer. https://doi.org/10.1007/978-3-319-09767-1_54.

Roberts, Adam, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck. 2018. "A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music." *arXiv,* arXiv: 1803.05428 [cs.LG].

Roberts, Richard M., and Roger J. Kreuz. 1994. "Why Do People Use Figurative Language?" *Psychological Science* 5 (3): 159–163. https://doi.org/10.1111/j.1467-9280.1994.tb00653.x.

Robertson, Judy, Andrew {de Quincey}, Tom Stapleford, and Geraint Wiggins. 1998. "Real-time music generation for a virtual environment" [in English]. In *ECAI 98 Workshop on AI/ALIFE and Entertainment,* edited by {Henri M } Prade. ISBN: 9780471984313.

Rocha de Azevedo Santos, Lucas, Carlos N. Silla Jr., and Marcelo Costa-Abreu. 2021. "A Methodology for Procedural Piano Music Composition with Mood Templates Using Genetic Algorithms." In *Proceedings of the 11th International Conference on Pattern Recognition Systems (ICPRS).*

Rockwell, Patricia. 2000. "Lower, slower, louder: Vocal cues of sarcasm." *Journal of Psycholinguistic Research* 29 (5): 483–495.

Russell, James A. 1980. "A Circumplex Model of Affect." *Journal of Personality and Social Psychology* 39 (6): 1161–1178.

———. 1983. "Pancultural Aspects of the Human Conceptual Organization of Emotions." *Journal of Personality and Social Psychology* 45 (6): 1281–1288.

Russell, James A. 2003. "Core Affect and the Psychological Construction of Emotion." *Psychological Review* 110 (1): 145–172.

Russell, James A., Anna Weiss, and Gerald A. Mendelsohn. 1989. "Affect Grid: A Single-Item Scale of Pleasure and Arousal." *Journal of Personality and Social Psychology* 57 (3): 493–502. https://doi.org/10.1037/0022-3514.57.3.493.

Russell, Stuart, and Peter Norvig. 2010. *Artificial Intelligence: A Modern Approach.* 3rd ed. Upper Saddle River, NJ: Prentice Hall.

Ryu, Seungbae. 2023. "Sentiment Synthesis: Immersive Virtual Theater with Real-Time Emotion and Sentiment Analysis." PhD diss., Virginia Polytechnic Institute and State University.

Saarimäki, H. 2021. "Cerebral Topographies of Perceived and Felt Emotions." *Trends in Cognitive Sciences* 25 (7): 459–471.

Saha, Anisha, Varsha Suresh, Timothy Hospedales, and Vera Demberg. 2025. *MUStReason: A Benchmark for Diagnosing Pragmatic Reasoning in Video-LMs for Multimodal Sarcasm Detection.* ArXiv:2510.23727 [cs.LG]. https://doi.org/10.48550/arxiv.2510.23727.

Sainz-de-Baranda Andujar, Clara, María Gutiérrez-Martín, José Miranda-Calero, María Blanco-Ruiz, and Celia López-Ongil. 2022. "Gender biases in the training methods of affective computing: Redesign and validation of the Self-Assessment Manikin in measuring emotions via audiovisual clips." *Frontiers in Psychology* 13:955530. https://doi.org/10.3389/fpsyg.2022.955530.

Saufnay, Sarah, Elodie Etienne, et al. 2025. "Speech Emotion Recognition for Public Speaking Training in Virtual Reality." In *AIXVR'26.* SER system designed for VR training contexts.

Schachter, Stanley, and Jerome E. Singer. 1962. "Cognitive, Social, and Physiological Determinants of Emotional State." *Psychological Review* 69 (5): 379–399. https://doi.org/10.1037/h0046234. https://doi.org/10.1037/h0046234.

Scharrer, Lisa, Ursula Christmann, and Maria Knoll. 2011. "Voice modulations in irony." *Journal of Pragmatics* 43:105–117.

Scherer, Klaus R. 1999. "Appraisal Theory." In *Handbook of Cognition and Emotion,* edited by Tim Dalgleish and Mick Power, 637–663. New York: Wiley.

———. 2003. "Vocal communication of emotion: A review of research paradigms." *Speech Communication* 40 (1-2): 227–256.

———. 2004. "Which Emotions Can be Induced by Music? What Are the Underlying Mechanisms?" *Journal of New Music Research* 33 (3): 239–251.

———. 2005. "What are emotions? And how can they be measured?" *Social Science Information* 44 (4): 693–727. https://doi.org/10.1177/0539018405058216.

Schlosberg, Harold. 1954. "Three Dimensions of Emotion." *Psychological Review* 61 (2): 81–88.

Schuller, Björn, Anton Batliner, Stefan Steidl, and Dino Seppi. 2011. "Recognising Realistic Emotions and Affect in Speech: State of the Art and Lessons Learnt from the First Challenge." *Speech Communication* 53 (9–10): 1062–1087.

Schuller, Björn, Stefan Steidl, and Anton Batliner. 2009. "The INTERSPEECH 2009 Emotion Challenge." *Proceedings of Interspeech.*

Schuller, Björn W., and Anton Batliner. 2014. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing.* John Wiley & Sons.

Scirea, M., J. Togelius, P. Eklund, and S. Risi. 2017. "Affective Evolutionary Music Composition with MetaCompose." *Genetic Programming and Evolvable Machines* 18:433–465.

Sharma, Gaurav, Jill Drury, and Gregory Ferrin. 2015. "Using Speech and Multimodal Interaction in Virtual Reality Environments." In *Proceedings of the IEEE Virtual Reality Conference,* 209–210.

Shaver, Phillip, Judith Schwartz, Donald Kirson, and Cary O'Connor. 1987. "Emotion Knowledge: Further Exploration of a Prototype Approach." *Journal of Personality and Social Psychology* 52 (6): 1061–1086.

Shrout, Patrick E., and Joseph L. Fleiss. 1979. "Intraclass correlations: Uses in assessing rater reliability." *Psychological Bulletin* 86 (2): 420–428. https://doi.org/10.1037/0033-2909.86.2.420.

Shum, Heung-Yeung, and Sing Bing Kang. 2000. "Review of Image-Based Rendering Techniques." *International Journal of Computer Vision* 36 (2): 151–170. https://doi.org/10.1023/A:1008197410864.

Singer, Uriel, Adam Polyak, Thomas Hayes, et al. 2022. "Make-A-Video: Text-to-Video Generation without Text-Video Data." *arXiv preprint arXiv:2209.14792.*

Slater, Mel, and Maria V. Sanchez-Vives. 2016. "Enhancing Our Lives with Immersive Virtual Reality." *Frontiers in Robotics and AI* 3:74. https://doi.org/10.3389/frobt.2016.00074. https://www.frontiersin.org/articles/10.3389/frobt.2016.00074/full.

Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. "Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank." *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*

Somarathna, Sanduni, Tomasz Bednarz, and Haifeng Wang. 2022. "Affective Virtual Reality: A Review of Emotion Elicitation and Recognition in Immersive Environments." *IEEE Transactions on Affective Computing* 13 (4): 1842–1860. https://doi.org/10.1109/TAFFC.2022.3181053.

Sra, Misha, Pattie Maes, Prashanth Vijayaraghavan, and Deb Roy. 2017. "Auris: Creating Affective Virtual Spaces from Music." In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology (VRST '17),* 26:1–26:11. ACM. ISBN: 978-1-4503-5548-3. https://doi.org/10.1145/3139131.3139139. https://dl.acm.org/doi/10.1145/3139131.3139139.

Stability AI. 2023. *Announcing Stable Audio, a product for music & sound generation.* https://stability.ai/news/stable-audio-using-ai-to-generate-music. Accessed 2026-01-02, September.

Suhaimi, Nurul Afiqah, James Mountstephens, and Jason Teo. 2020. "Emotion Recognition in Immersive 360-Degree Video Environments Using Physiological Signals." *IEEE Access* 8:123400–123415. https://doi.org/10.1109/ACCESS.2020.XXXXXXX.

Szenberg, Michael, and Lall Ramrattan. 2014. "Definitions of Irony." In *Economic Ironies Throughout History,* 9–23. New York: Palgrave Macmillan. https://doi.org/10.1057/9781137450821_2.

Taboada, Maite, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. "Lexicon-Based Methods for Sentiment Analysis." *Computational Linguistics* 37 (2).

Tan, Hao, and Dorien Herremans. 2020. "Music FaderNets: Controllable Music Generation Based on High-Level Features via Low-Level Feature Modelling." *arXiv,* arXiv: 2007 . 15474 [cs.SD].

Tang, Duyu, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016. "Effective LSTMs for Target-Dependent Sentiment Classification." *Proceedings of the International Conference on Computational Linguistics.*

Tencent Hunyuan Team. 2024. *Hunyuan World: Interactive 3D World Generation Model.* Technical report and open-source release.

Tepperman, Joseph, David Traum, and Shrikanth Narayanan. 2006. ""Yeah right": Sarcasm recognition for spoken dialogue systems." *Proceedings of Interspeech,* 1838–1841.

Thayer, Robert E. 1989. *The Biopsychology of Mood and Arousal.* Oxford University Press.

Trigeorgis, George, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A. Nicolaou, Björn Schuller, and Stefanos Zafeiriou. 2016. "Adieu Features? End-to-End Speech Emotion Recognition Using a Deep Convolutional Recurrent Network." In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

Turney, Peter D. 2002. "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews." *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL),* 417–424. https://doi.org/10.3115/1073083.1073153. https://aclanthology.org/P02-1053/.

Ubisoft. 2017. *Star Trek: Bridge Crew.* Commercial VR Game Release; includes integrated voice command system using IBM Watson NLP.

Van Hee, Cynthia, Els Lefever, and Véronique Hoste. 2018. "SemEval-2018 Task 3: Irony Detection in English Tweets." *Proceedings of the International Workshop on Semantic Evaluation.*

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." In *Advances in Neural Information Processing Systems.*

Ververidis, Dimitrios, and Constantine Kotropoulos. 2006. "Emotional Speech Recognition: Resources, Features, and Methods." *Speech Communication* 48 (9): 1162–1181.

Vogt, Thurid, and Elisabeth André. 2008. "Improving Automatic Emotion Recognition from Speech via Gender Differentiation." In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC),* 1129–1134.

Vogt, Thurid, Yiorgos Chryssanthou, Amaryllis Raouzaiou, and Elisabeth André. 2009. "CALLAS: How to Use Multimodal Emotion Analysis in Interactive Art." In *Proceedings of the International Conference on Technologies for Interactive Digital Storytelling and Entertainment (TIDSE),* 83–94. Springer.

Wallis, Isaac. 2008. "Computer Generating Emotional Music: The Design of an Affective Music Generator." Rule-based affective music generation grounded in music-theoretical parameters and valence–arousal evaluation. PhD diss., Arizona State University. https://asu.elsevierpure.com/en/publications/computer-generating-emotional-music-the-design-of-an-affective-mu/.

Wang, Yingzhi, Abdelmoumene Boumadane, and Abdelwahab Heba. 2022. "A Fine-tuned Wav2vec 2.0/HuBERT Benchmark For Speech Emotion Recognition, Speaker Verification and Spoken Language Understanding." In *Proc. Interspeech 2022,* 1546–1550. https://doi.org/10.21437/Interspeech.2022-10088. https://arxiv.org/abs/2111.02735.

Watson, David, and Lee Anna Clark. 1994. *The PANAS-X: Manual for the Positive and Negative Affect Schedule – Expanded Form.* Technical report. University of Iowa.

Watson, David, Lee Anna Clark, and Auke Tellegen. 1988. "Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales." *Journal of Personality and Social Psychology* 54 (6): 1063–1070.

Watson, David, and Auke Tellegen. 1985. "Toward a Consensual Structure of Mood." *Psychological Bulletin* 98 (2): 219–235.

Wedin, Lars. 1972. "A Multidimensional Study of Perceptual-Emotional Qualities in Music." *Scandinavian Journal of Psychology* 13:241–254.

Williams, Duncan, Alexis Kirke, Eduardo R. Miranda, Etienne Roesch, Ian Daly, and Slawomir J. Nasuto. 2015. "Investigating Affect in Algorithmic Composition Systems." *Psychology of Music* 43 (6): 831–854. https://doi.org/10.1177/0305735614543282.

Wilson, Deirdre, and Dan Sperber. 2012. "Explaining Irony." In *Meaning and Relevance,* edited by Deirdre Wilson and Dan Sperber, 123–145. Cambridge: Cambridge University Press.

Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann. 2005. "Recognizing contextual polarity in phrase-level sentiment analysis." In *Proceedings of HLT/EMNLP.*

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, et al. 2020. "Transformers: State-of-the-Art Natural Language Processing." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations,* 38–45. Online: Association for Computational Linguistics, October. https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Wu, Shih-Yu, Tiago H. Falk, and William-Y. Chan. 2014. "Automatic Instrument-Based Affect Recognition in Music Using Timbre Features." In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR),* 513–518.

Wundt, Wilhelm. 1905. *Grundzüge der physiologischen Psychologie.* Wilhelm Engelmann.

Xu, Bin, Shuang Wang, and Xia Li. 2010. "An Emotional Harmony Generation System." In *Proceedings of the IEEE Congress on Evolutionary Computation (CEC),* 1–6.

Yu, Wenmeng, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. "CH-SIMS: A Chinese Multimodal Sentiment Analysis Dataset with Fine-grained Annotation of Modality." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020),* 3718–3727. Online.

Zadeh, Amir, Paul Pu Liang, Jaret Vanbriesen, Soujanya Poria, Emily Tong, Erik Chang, and Louis-Philippe Morency. 2018. "Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Benchmark." *Proceedings of the Annual Meeting of the Association for Computational Linguistics.*

Zadeh, Amir, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. "MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos." *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*

Zhang, Q., X. Chen, Q. Zhan, T. Yang, and S. Xia. 2017. "Respiration-based emotion recognition with deep learning." *Computers & Industry* 92:84–90.

Zhang, X. 2025. "Visualization of Poetry Recitation Emotions in VR." In *ACM/IEEE VR or similar conference proceedings.*

Zhao, Kun, Siqi Li, Juanjuan Cai, Hui Wang, and Jingling Wang. 2019. "An Emotional Symbolic Music Generation System based on LSTM Networks." In *2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC),* 2039–2043. IEEE. https://doi.org/10.1109/ITNEC.2019.8729266.

Zheng, Simeng Gu, Yu Lei, Shanshan Lu, Wei Wang, Yang Li, and Fei Wang. 2016. "Safety Needs Mediate Stressful Events Induced Mental Disorders." *Neural Plasticity* 2016:1–6. https://doi.org/10.1155/2016/8058093.

Zheng, Kaitong, Ruijie Meng, Chengshi Zheng, Xiaodong Li, Jinqiu Sang, Juanjuan Cai, and Jie Wang. 2022. "EmotionBox: A Music-Element-Driven Emotional Music Generation System Based on Music Psychology." *Frontiers in Psychology* 13:841926. https://doi.org/10.3389/fpsyg.2022.841926.

Zhu, Hong, Shangfei Wang, and Zhaohui Wang. 2008. "Emotional Music Generation Using Interactive Genetic Algorithm." In *Proceedings of the International Conference on Computer Science and Software Engineering,* 1:345–348.

Zuckerman, Marvin, Bernard Lubin, L. Vogel, and E. Valerius. 1964. "Measurement of Experimental Affects." *Journal of Consulting Psychology* 28 (5): 418–425.

Appendix A

Documentation and Repository Structure

This appendix provides comprehensive documentation of the computational frameworks and creative iterations developed for the research project *The Ironic Machine[s]*. All source code, experimental datasets, and interactive builds are hosted in the following version-controlled repository: https://github.com/jfforero/The-Ironic-Machines/

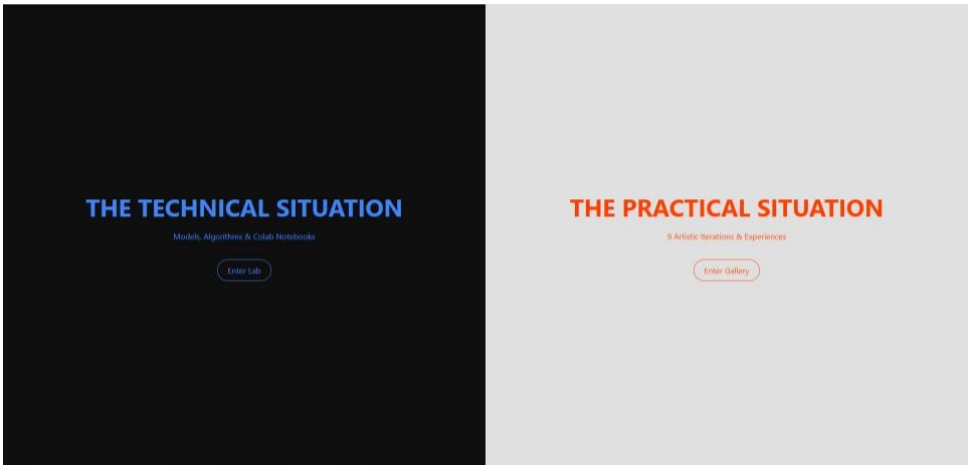


Figure A.1: Overview of the GitHub repository structure and technical situations.

The repository is organized to mirror the structural logic of the project's official website: https://www.emotional-machines.com. While the website served as a real-time research blog and process journal during the doctoral candidacy, the GitHub repository serves as the definitive archival source for the thesis.

It is important to note that while the web-based documentation provides a high-fidelity interactive experience, the repository ensures the long-term reproducibility of the research. This dual-layered approach balances the dynamic nature of practice-based research with the rigorous archival requirements of a doctoral thesis in Computer Engineering.

A.1 Technical Situation: Scientific Methodology

The analytical backbone of this research is implemented through a series of modular Python-based pipelines. These modules handle the bimodal extraction of affect and the subsequent generative mapping for virtual environments.

Table A.1: Summary of Technical Implementation Notebooks

Component	**Functionality**	**Repository Access**
Sentiment Analysis	BERT/LSTM polarity detection for environmental modulation.	View on Colab
Speech Emotion Recognition	MFCC and eGeMAPS extraction for acoustic arousal/valence.	View on Colab
Irony Analysis	Statistical divergence modeling between semantic and acoustic vectors.	View on Colab
Affective Music Generation	Computer-based composition conditioned on V-A coordinates.	View on Colab
Affective Environment Generation	Text guided equirectangular panoramas generation	View on Colab

A.2 Practical Situation: Iterative Development

The research followed a practice-based methodology resulting in nine iterations. These iterations document the evolution from simple audio-visual synthesis to complex, irony-aware Affective Virtual Environments (AVE).

1. **Iteration 01: What Stirs Within** – Initial web-based affective audio synthesis and voice-to-space mapping. *Access: [GitHub]*

2. **Iteration 02: A Máquina Sensível** – Real-time voice interaction within a Unity3D environment using procedural geometry. *Build: [Itch.io]*

3. **Iteration 03: Abandoned Memories** – Neural style transfer guided by speech input to evoke nostalgia. *Documentation: [Web Page]*

4. **Iteration 04: Veintitantos Vientos** – Spatial reconfiguration using 24 equirectangular panoramas triggered by poetic utterances. *Documentation: [Web Page]*

5. **Iteration 05: En train d'oublier** – Web-art exploration of memory decay in the Salar de Uyuni. *Access: [WordPress]*

6. **Iteration 06: Terra Australis Ignota** – Speculative emotional cartography, exhibited at Florence XR 2024. *Access: [WordPress]*

7. **Iteration 07: The What Space** – VR environment focused on the prosodic variability of the word "What". *Access: [WordPress]*

8. **Iteration 08: The Ironic Machine[s]** – Primary implementation of the bimodal irony framework contrasting sarcasm and kindness. *Interactive: [Project Site]*

9. **Iteration 09: Bello** – Final synthesis mapping recited poetry into 360° evolving landscapes at Igreja da Sé. *Documentation: [Web Page]*

A.3 Visual Documentation and Artifacts

The following figure provides a chronological record of the nine iterations, capturing the real-time modulation of geometry, lighting, and environmental textures based on bimodal affective input.

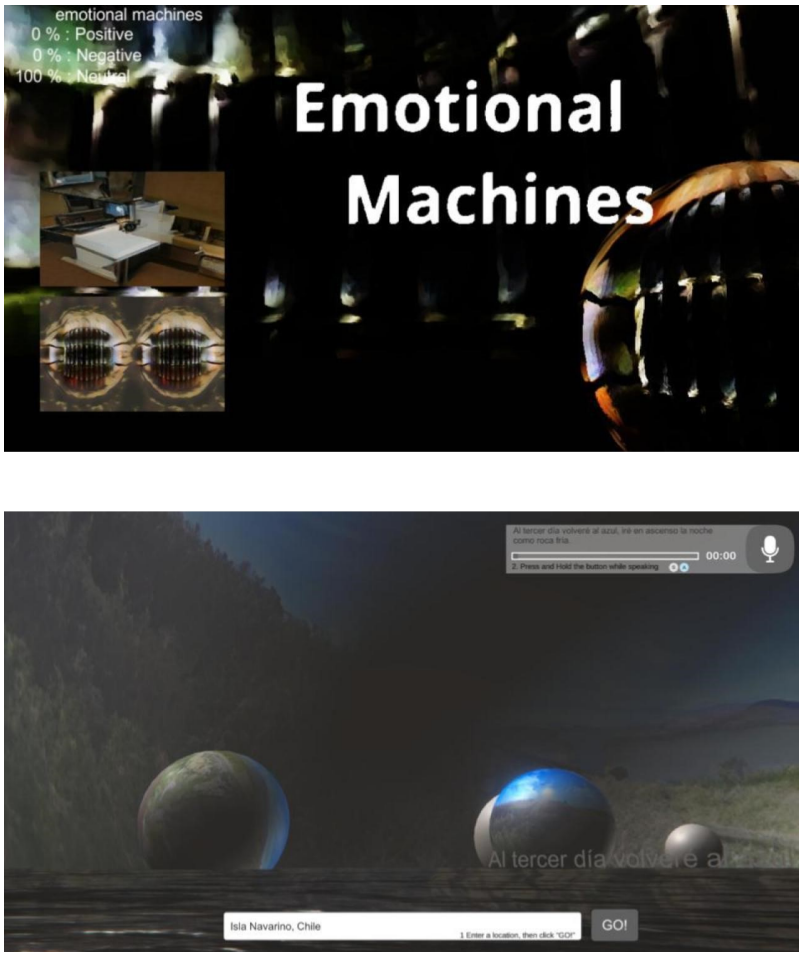


Figure A.2: Visual documentation of the iterative process, showcasing screenshots from the primary Affective Virtual Environments (AVE).

A comprehensive image gallery and technical log of screenshots taken during the developmental process are available via the following digital repository: [Technical Image Gallery]

A.4 Poe++

The research incorporates several poetical projects that explore the intersection of machine learning, affective linguistics, and creative expression. These artifacts serve as the qualitative output for the "Rim-Bot" persona and the bimodal irony framework.

- **Proyecto Poético: Rim-Bot vuelve a casa** – A central creative pillar of the research, investigating the machine's "return" to emotional expression. *Access: [Rim-Bot Project]*
- **Audio Poemas Edípicos** – A collection of sonified poems exploring vocal prosody and emotional weight. *Access: [Audio Poems]*
- **Esquizo-poes[IA]** – An experimental exploration of AI-generated poetic structures and "schizo-analytic" linguistic patterns. *Access: [Esquizo-Poetry]*

Figure A.3: Vous êtes fou, monsieur Rim-bot?