Desiring Machines and Affective Virtual Environments

Jorge Forero[1,2,3](✉), Gilberto Bernardes[1,2], and Mónica Mendes[3]

[1] Faculty of Engineering, University of Porto, 4200-465 Porto, Portugal
jfforero@ludique.cl
[2] INESC TEC, 4200-465 Porto, Portugal
[3] ITI-LARSyS, Faculdade de Belas Artes, Universidade de Lisboa, Lisbon, Portugal

Abstract. Language is closely related to how we perceive ourselves and signify our reality. In this scope, we created **Desiring Machines**, an interactive media art project that allows the experience of affective virtual environments adopting speech emotion recognition as the leading input source. Participants can share their emotions by speaking, singing, reciting poetry, or making any vocal sounds to generate virtual environments on the run. Our contribution combines two machine learning models. We propose a long-short term memory and a convolutional neural network to predict four main emotional categories from high-level semantic and low-level paralinguistic acoustic features. Predicted emotions are mapped to audiovisual representations by an end-to-end process encoding emotion in virtual environments. We use a generative model of chord progressions to transfer speech emotion into music based on the tonal interval space. Also, we implement a generative adversarial network to synthesize an image from the transcribed speech-to-text. The generated visuals are used as the style image in the style-transfer process onto an equirectangular projection of a spherical panorama selected for each emotional category. The result is an immersive virtual space encapsulating emotions in spheres disposed into a 3D environment. Users can create new affective representations or interact with other previously encoded instances (This ArtsIT publication is an extended version of the earlier abstract presented at the ACM MM22 [1]).

Keywords: Affective Computing · Speech Emotion Recognition · Intelligent Virtual Environments · Virtual Reality · Tonal Interval Space · Machine Learning

1 Introduction

The Nobel Prize laureate Richard Feynman was asked in a 1985 lecture if machines will ever think like humans and if they will surpass our intelligence. Feynman claimed that machines would never think like humans and argued that we should first attempt to understand what intelligence is. Minsky proposes that emotions are ways to think [2]. In this context, Picard is credited for the branch of computer science at MIT known as affective computing. She argues that "affective computers should not only provide better

A. L. Brooks (Ed.): ArtsIT 2022, LNICST 479, pp. 405–414, 2023.
https://doi.org/10.1007/978-3-031-28993-4_28

performance in assisting humans, but also they might enhance computers' abilities to make decisions" [3].

Emotion recognition is a research field that attracts transversal interest in various disciplines of knowledge, such as linguistics [4], phonetics [5], psychology [6, 7], psychiatry [8], and computer science [9–11], among other fields of study.

From an overly-simplistic interpretation of Mehrabian's communication theory, 55% of the message is contained in facial expressions. In comparison, 38% is related to paralinguistic information (the way words are said), and only 7% of the message pertains to the semantic sense of words expressed (what is said) [12].

Computer vision technologies for face recognition raise significant limitations in virtual reality (VR), where the adoption of headsets usually hinders parts of the user's face. In this sense, language provides a natural interaction that benefits the emotion recognition process. Affective systems in VR involve developing at least two components: an emotion detection technique and a virtual environment generator [13].

The contributions of our research are of technical, practical, and critical nature. The main technical novelty is the combination of two machine learning models to predict emotions from semantic and acoustic features. Furthermore, we propose a practical end-to-end pipeline to create visual representations of the users' emotional predictions. The artistic proposal contributes to the discussion on how language shapes our realities. Adopting a post-structuralist Deleuzian perspective, we recognize the need to think about the construction of spaces stripped of all lack and provided by desire as a creative tool. The aim is to promote an experience where users can reflect on the relation between speech, emotion, and virtual environments.

The remainder of this paper is organized to answer the artwork's technical, practical, and critical situation as follows: Sect. 2 (the technical situation) shows an overview of the evolution of intelligent environments using automatic speech recognition. Section 3 (the practical situation) exposes our proposal for the recognition model and the generative system. Finally, Sect. 4 (the critical situation) briefly introduces the Deleuzian idea of desire as a creative tool.

2 Intelligent Virtual Environments: The Technical Situation

Aylett and Cavazza first described the intersection between artificial intelligence and virtual environments (VEs) as Intelligent Virtual Environments (IVEs) [14]. There has been an increasingly growing interest in developing IVEs in many fields. Remarkably, there has been a continuous interest in automatic speech recognition technologies for virtual reality applications [15, 16]. As far as we investigate from public resources, the first use of speech recognition and virtual reality can be traced to a 1995 project by the US Navy Research Laboratory [17]. In 1996, Rose and Wilhelms described an interface for designing 3D scenes in VEs using ASR [18]. In their system, speech commands replaced traditional interfaces to explore creative ideas and build 3D scenes quickly. In 2002, Levin and Lieberman presented an augmented-reality speech-visualization system named The Hidden Worlds of Noise and Voice [19]. The project allows users to visualize the voices captured by the system, which are visualized in the form of animated graphics emerging from the mouths of the users while speaking. The graphics representing these expressions assume various shapes and behaviors associated with basic characteristics of the vocalist's volume, pitch, and timbre. Maes and her team at MIT present a system named Auris that automatically generates VR environments from music [20]. As input sources, they proposed using songs (audio and lyrics) to produce a VR world that encapsulates the mood and content of the song in the space design. Using creative licenses, they transform the generated virtual landscapes into psychedelic and surreal places by pre-processing textures through the DeepDream neural network that they subsequently apply to 3D objects. In 2021, Pinilla et al. suggested an association between features of audio-visual elements and affective states. They propose guiding cues to develop virtual environments that automatically create visual representations of users' affective states, analyzing their electrophysiological activity [13].

3 Our Proposal: The Practical Situation

Desiring Machines is a digital media project that proposes an experience where participants can share their emotions using their voices to generate audiovisual representations in virtual reality. Our research proposes using two machine learning (ML) models to predict speech emotions from semantic and acoustic features. As shown in Fig. 1, sentimental analysis and speech emotion predictions are mapped into Affective Virtual Environments (AVE) represented by spheres encoding participants' experiences.

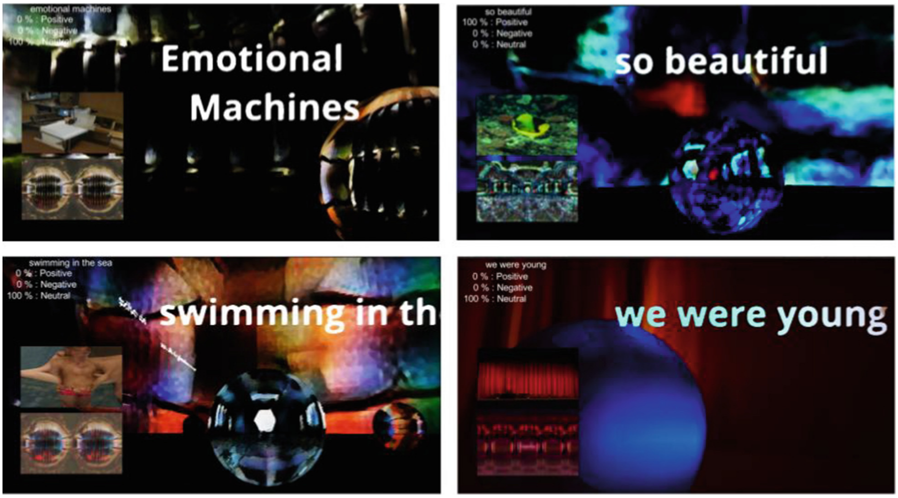

Fig. 1. Examples of four virtual environments created using speech emotion recognition.

3.1 Machine Learning Architecture

Figure 2 shows a general architecture of our affective virtual environment system. It includes two main modules responsible for the speech emotion recognition process and the virtual environment generation.

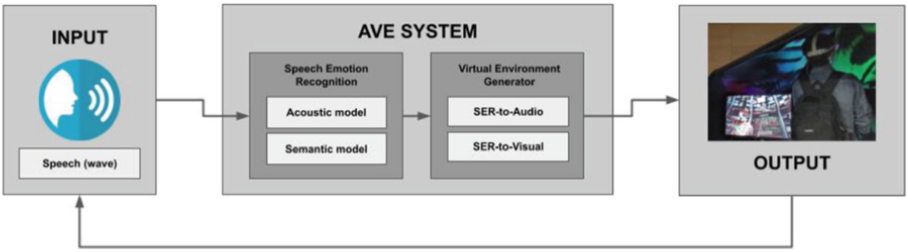

Fig. 2. General Architecture of the affective virtual environment system.

3.2 Speech Emotion Recognition System

We included the speech-to-text procedure through Microsoft's automatic speech recognition application programming interface (API). A sentimental analysis of the transcribed words or sentences evaluates the emotional polarity. We implement a long-short-term memory (LSTM) network architecture to get context information (Fig. 3). To train the network, we used two databases containing labeled poetry and film reviews; Poem Sentiment is a dataset of verses randomly picked from the Gutenberg project [21]. Each sample was labeled as negative, no impact, positive, or mixed (both negative and positive). We also included the Internet Movie Database (IMDb). IMDb is a large Movie review dataset containing 50.000 annotated files with positive and negative labels [22]. Text strings are pre-processed with the python Natural Language Toolkit (NLTK), substituting regular expressions, removing stop words, and using stemming and lemmatization. We used Keras Tokenizer to vectorize and convert text into Sequences with a length of 1884 tokens. The network was trained to minimize the categorical Crossentropy loss function for four epochs with an accuracy of 88% on a validation dataset of 150 new text files.

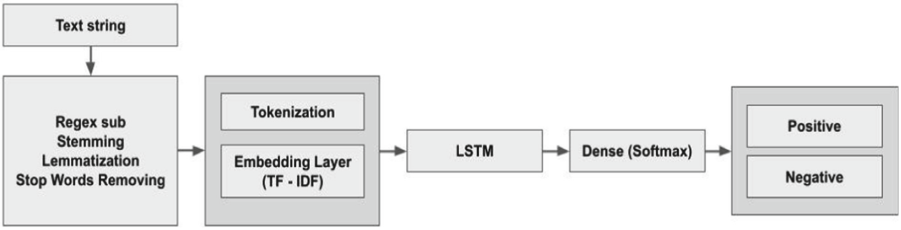

Fig. 3. Semantic Model for sentiment analysis of a text string.

We propose a convolutional neural network (CNN) architecture for the paralinguistic acoustic counterpart. The resulting model allows us to predict emotions from the audio clips recorded and temporarily saved in the CPU. We have used three different datasets to train this model: We incorporated 1440 audio files from The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [23]. Speeches were recorded by 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprised, and disgust. Each statement has been recorded at two levels of emotional intensity (normal and strong), with an additional neutral expression. We also added the VIVAE corpus of non-verbal audio speeches in English. The Variably Intense Vocalizations of Affect and Emotion [24] comprises 1085 audio files of eleven young females. Each audio is labeled on his filename under six affective states (anger, fear, pain, achievement/triumph, positive surprise, and sexual pleasure), each with four intensity levels (low, moderate, strong, and peak). Finally, we included CREMA-D [25], an audiovisual acted database composed of 7,442 files labeled under six emotional categories (happy, sad, anger, fear, disgust, and neutral). All three Databases were merged into two categories (positive and negative audio clips). In total, we grouped 1821 positive and 5619 negative audio files. To expand our merged database, we augmented audio, adding white noise, shifting time and pitch, and stretching the original audio signals. A combination of spectral acoustics features was extracted from all the audio clips to generate the input vectors. These vectors were passed through four one-dimensional CNN that output the Feature maps using the Relu activation functions (Fig. 4). We considered; Mel spectrogram (mean for each of 128 bands), MFCC (64 coefficients), and the chroma vector (12 features) for each audio file. We split the data randomly in a 4:1 ratio into training and validation sets. The network was trained by minimizing the categorical Cross Entropy loss for 100 epochs in batches of 64 files with an accuracy of 81% on the testing dataset.

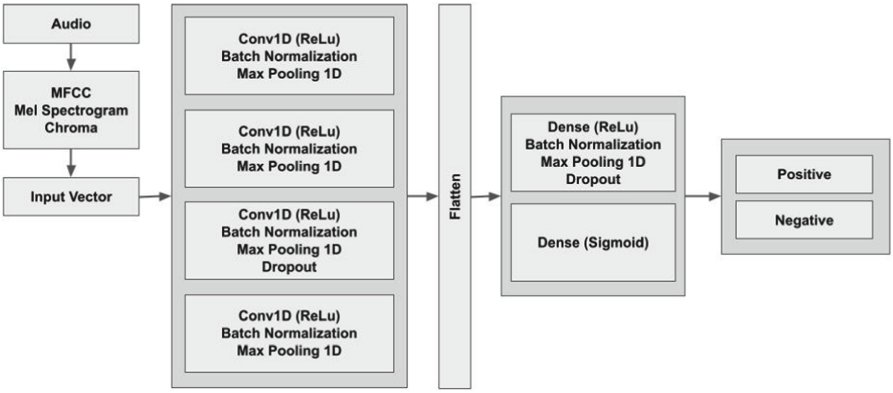

Fig. 4. Acoustic Model for speech emotion recognition.

The semantic sentiment analysis and the paralinguistic acoustic machine learning models are merged to produce one of four predicted emotion categories shown in Table 1. Predictions carried out are then encoded into a 3D virtual environment presented through a virtual reality headset.

Table 1. Possible emotion predictions in the AVE system.

Semantic	Acoustic	Result
Positive	Positive	Positive
Positive	Negative	Irony −
Negative	Positive	Irony +
Negative	Negative	Negative

We define semantic-acoustic contradictions as ironies. A negative irony would involve a positive semantic statement and a negative acoustic prediction. A positive irony is composed of a positive sentiment analysis and a negative acoustic classification.

3.3 Virtual Environment Generation

To transfer speech emotion to audio, we implement a ruled-based model to create chord progressions based on consonance and harmonic dispersion (i.e., distance-to-key) in the Tonal Interval Space [26]. We use the Conchord framework to map emotional predictions (Fig. 5) into the acoustical virtual environment.

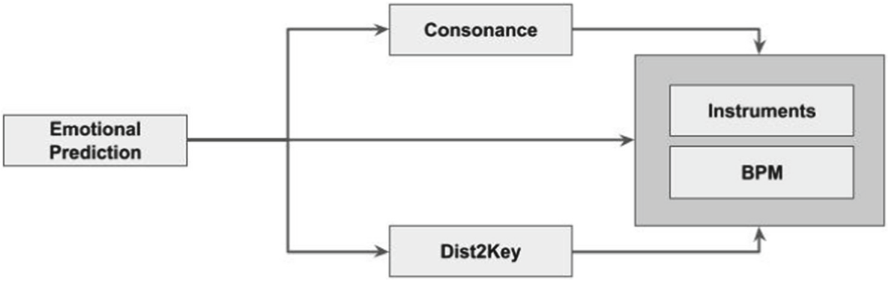

Fig. 5. Speech emotion recognition to audio.

Positive results are mapped to a high level of consonance and a low level of dispersion. Negative results are mapped to a low level of consonance and a high-level dispersion. Progressions are connected to a set of instruments selected for each emotion. Ironies are mapped in between according to the two parameters mentioned above.

We built the visual representation adopting a text-to-image API provided by DeepAI, prompted with the text transcribed from speeches uttered. The resulting image is scaled and used as a style source transferred over an equirectangular panorama chosen for each emotional category defined (Fig. 6).

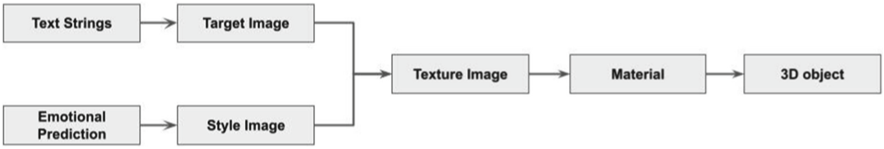

Fig. 6. Speech emotion recognition to visuals.

The artistic result is an immersive virtual space that encodes emotions in spheres disposed into a virtual reality environment. Participants can generate new affective representations or interact with others previously created using joysticks (Fig. 7).

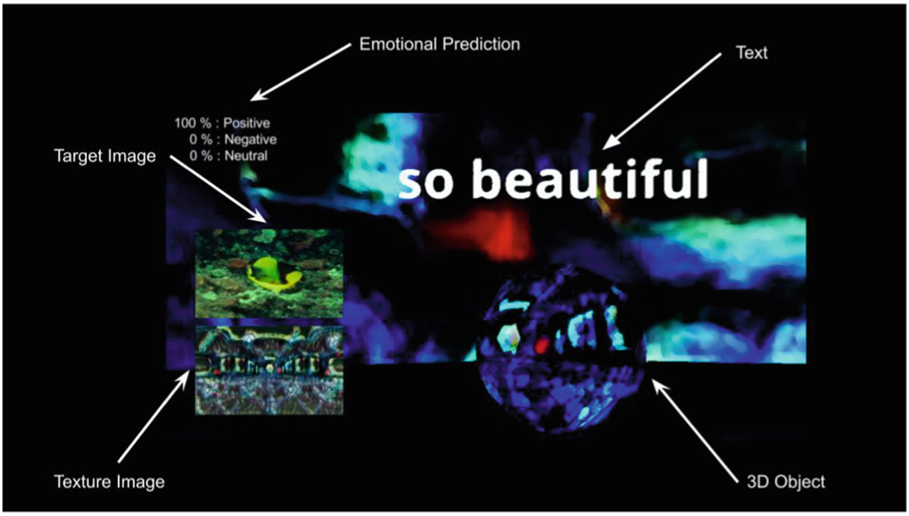

Fig. 7. Affective virtual environments generation.

3.4 The User Interface

As shown in Fig. 8, the user interface comprises a head-mounted display and two joysticks. The HMD is equipped with a lens, headphones, and a microphone. The HMD is connected to a Central Unit (CPU) where speech is processed. The UI also proposes publicly sharing a stream from the user's point of view. When there is no activity, the stream shows preselected recorded performances.

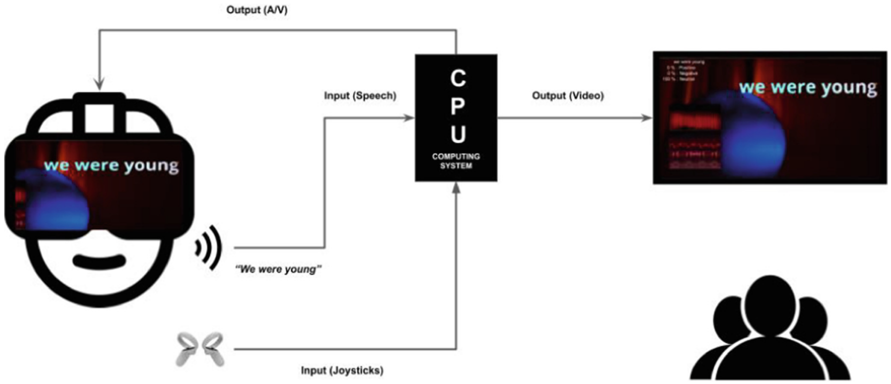

Fig. 8. User Interface general configuration.

4 Emotional Machines Are Desiring Machines (The Critical Situation)

For Ferdinand de Saussure, while speaking is an individual activity, language is a social manifestation of speech. In his conception, language is a set of signs that evolves from speech activity. As he exposes, these signs are constituted by signifiers and signified. For the author, while the signifier is the acoustic impression of a concept, the signified is its mental correlation. In Jacques Lacan, the subject uses signifiers to perceive the world, everything that surrounds it, and itself; it enters the world structured by language. For him, humanity does not determine language; on the contrary, language defines (and constrains) the world and humanity. Lacan relates reality with two central concepts of psychoanalytic theory, the ghost and the desire [27]. Lacan is heir to the Freudian tradition in that the origin of desire would arise from rediscovering an original (and lost) experience of satisfaction, which appears in the relationship with the object of desire. In this sense, desire results from a lack, and emotions are repressed by the signifier in the speaking being.

A different perspective is adopted by Deleuze and Guattari, for whom desire does not seek to objectify itself from a lack that must be supplied. Deleuze and Guattari explicitly question this psychoanalytic and structuralist conception of desire. What constitutes the central theme of Anti-Oedipus is that, for these authors, desire is a factory that constantly produces. The objections of The Anti-Oedipus to structural linguistics are fundamentally directed against the primacy granted to the signifier since it performs an oppressive social function. In the Anti-Oedipus, the genealogy of lack is relegated to the detriment of the post-structuralist conception of flow [28]. Thus they propose an alternative ontology of desire built from a lack by giving it a creative capacity.

In "*What is Philosophy?*" Deleuze describes how this creative process produces something new, namely '*affects*' and '*percepts*'. Artistic creativity generates new affects and percepts and combines them into blocks of sensation [29]. These blocks then form the basis of our experienced reality. For the French authors Deleuze and Guattari, desiring machines are the site of that production.

5 Conclusion

Language and emotions are intimately related. From a Deleuzian perspective, language has to release itself from a foundational lack and transit to a creative approach of desire. In this sense, we find in poetry a way to overcome language structure by following the rules of imagination.

In this research project in digital media arts, we propose a theoretical, practical, and critical situation to define and create affective virtual environments that can be generated or modulated by predicting emotion in speeches.

Our proposal seeks to contribute to the technical domain by combining two machine-learning models that account for speech's semantic and acoustic dimensions.

Our intentionally reductionist ML models regarding the polar selection of the contents lead us to ask ourselves about the affective spaces of representation from a more complex perspective. Which are the best speech features to predict emotions for those representation spaces?

The environments created with the developed system allow us to experiment with the use of languages according to different linguistic perspectives and thus, in a certain way, return the organs to their mother cell where all new creations are possible.

Acknowledgments. This research was partially funded by the project "Experimentation in music in Portuguese culture: History, contexts and practices in the 20th and 21st centuries" (POCI-01-0145-FEDER-031380) co-funded by the European Union through the Operational Program Competitiveness and Internationalization, in its ERDF component, and by national funds, through the Portuguese Foundation for Science and Technology (FCT). It was also partially funded by FCT in the scope of the projects: UIDB/11918/2022 and UIDB/50009/2020.

References

1. Forero, J., Bernardes, G., Mendes, M.: Emotional machines: toward affective virtual environments. In Proceedings of the 30th ACM International Conference on Multimedia (MM 2022), pp. 7237–7238. Association for Computing Machinery, New York (2022). https://doi.org/10.1145/3503161.3549973
2. Minsky, M.: The Emotion Machine. Simon & Schuster (2006)
3. Picard, R.W.: Affective computing. M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 321 (1995)
4. Kitayama, S., Markus, H.R.: Emotion and culture: Empirical studies of mutual influence. Am. Psychol. Assoc. 1–19 (1994)
5. Roach, P.: Techniques for the phonetic description of emotional speech. In: Proceedings of the ISCA Workshop on Speech and Emotion (2000)
6. Russell, J.: A circumplex model of affect. J. Pers. Soc. Psychol. **39**(6), 1161 (1980)
7. Russell, J.: How shall an emotion be called? (1997)
8. Scherer, K.: What are emotions? And how can they be measured? Soc. Sci. Inf. **44**, 695–729 (2005)
9. Cowie, R., et al.: Emotion recognition in human-computer interaction. IEEE Signal Process. Mag. **18**, 32–80 (2001)
10. Oudeyer, P.-Y.: Novel useful features and algorithms for the recognition of emotions in human speech. In: International Conference on Speech Prosody 2002 (2002)

11. Petrushin, V.: Emotion in speech: recognition and application to call centers. In: Proceedings of Artificial Neural Networks in Engineering (2000)
12. Mehrabian, A., Wiener, M.: Decoding inconsistent communications. J. Pers. Soc. Psychol. **6**, 109–114 (1967)
13. Pinilla, A., Garcia, A., Raffe, W., Voigt-Antons, J.-N., Spang, R., Müller, S.: Affective visualization in virtual reality: an integrative review. Front. Virtual Reality **2**, 630731 (2021)
14. Aylett, R., Cavazza, M.: Intelligent virtual environments – a state-of-the-art report. In: Proceedings of the Eurographics Workshop in Manchester UK (2001)
15. Karlgren, J., Bretan, N., Jonsson, L.: Interaction models, reference, and interactivity for speech interfaces to virtual environments. In: Göbel, M. (ed.) Virtual Environments 1995. Eurographics, pp. 149–159. Springer, Vienna (1995). https://doi.org/10.1007/978-3-7091-9433-1_13
16. Kamath, R., Kamat, R.: Development of an intelligent virtual environment for augmenting natural language processing in virtual reality systems. Int. J. Emerg. Trends Technol. Comput. Sci. (IJETTCS) **2**, 198–203 (2013)
17. Everett, S., Wauchope, K., Pérez, M.: A natural language interface for virtual reality systems (1995)
18. Clay, S.R., Wilhelms, J.: Put: language-based interactive manipulation of objects. IEEE Comput. Graph. Appl. **16**, 31–39 (1996)
19. Levin, G., Lieberman, Z.: In-situ speech visualization in real-time interactive installation and performance (2004)
20. Sra, M., Maes, P., Vijayaraghavan, P., Roy, D.: Auris: creating affective virtual spaces from music. In: Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology (VRST 2017), pp. 1–11 (2017)
21. Sheng, E., Uthus, D.: Investigating societal biases in a poetry composition system. arXiv (2020)
22. Maas, A., Daly, R., Pham, P., Huan, D., Ng, A., Potts, C.: Learning word vectors for sentiment analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (2011)
23. Livingstone, S.R., Russo, F.A.: The ryerson audio-visual database of emotional speech and song (RAVDESS). Zenodo **13** (2018)
24. Holz, N., Larrouy-Maestri, P., Poeppel, D.: The variably intense vocalizations of affect and emotion (VIVAE) corpus prompts new perspectives on nonspeech perception. Emotion **22**, 213 (2022)
25. Cao, H., Cooper, D.G., Keutmann, M.K., Gur, R.C., Nenkova, A., Verma, R.: CREMA-D: crowd-sourced emotional multimodal actors dataset. IEEE Trans. Affect. Comput. **5**, 377–390 (2014)
26. Bernardes, G., Cocharro, D., Guedes, C., Davies, M.E.P.: Conchord: an application for generating musical harmony by navigating in the tonal interval space. In: Kronland-Martinet, R., Aramaki, M., Ystad, S. (eds.) CMMR 2015. LNCS, vol. 9617, pp. 243–260. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46282-0_15
27. Lacan, J.: Écrits, A Selection. Alan Sheridan (1977)
28. Deleuze, G.: L'Île déserte et autres textes: textes et entretiens 1953–1974. In: Lapoujade, D. (ed.) Les Éditions de Minuit (2002)
29. Deleuze, G., Guattari, F.: Qu'est-ce que la philosophie? N.p.: Editions de Minuit (1991)