Not I: The Voice, Identity and the Epistemic Mirage of Machine Learning

Martin Disley
m.disley@ed.ac.uk
Institute for Design Informatics, University of Edinburgh
Edinburgh, UK

Murad Khan
murad.khan@arts.ac.uk
Creative Computing Institute, University of the Arts London
London, UK

Not I is an essay film and multichannel installation that explores the problematics of contemporary vocal profiling technology. It challenges the impulse to apply statistical learning techniques to polymorphous features of human expression for the purpose of speculative reconstruction exhibited in data science. Building upon the work of scholars across sound studies and vocal studies, *Not I* troubles the assumptions behind attempts to distil a one-to-one mapping of voice and identity. The film centres on an investigation into, and attack on, Speech2Face, a machine learning model which attempts to generate an image of the face of a speaker based solely on a recording of their voice. We leverage the affective qualities of moving image work to present this investigation as a form of experiential critique, forcing the viewer into an affective scenario that unsettles their existing heuristics used to infer speaker identity from vocal perception.

Introduction

> *and not alone the lips ... the cheeks ... the jaws ... the whole face .... all those ... what? (Beckett 1972)*

Not I is an essay film and installation produced by the creative research studio Unit Test. Through an adversarial engagement with 'Speech-2Face' (Oh et al. 2019), a machine learning model which attempts to generate an image of the face of a speaker based solely on a recording

Keywords Essay Film, Deep Learning, Vocal Forensics, Adversarial Attack, Investigative Aesthetics.
DOI 10.34626/2024_xcoax_026

of their voice *Not I* explores the problematics at the heart of contemporary vocal forensics' use of statistical learning techniques, developing a form of investigative aesthetics which seeks to open up the epistemic assumptions that ground the development of these socio-technical objects through a 'counterculture' of machine learning (McQuillan 2018).

In their overview of the use of machine methods for speaker identification and recognition in the 20th Century, Xiochang Li and Mara Mills note the introduction of 'vocal portraits' into the criminal archives of police departments across Europe and the United States, where these auditory impressions of criminality were deployed to speak to the character of the individual" (Li and Mills 2019). Much as Francis Galton's composite photography (Galton 1879) sought to surface a typology of criminality through the averaging of criminal faces, its phonographic equivalent also endeavoured to represent the gradient of social deviance through a new form of criminology. This turn to 'the probabilistic' – a focus upon estimates, approximations and intuitions about behavioural features, rather than an analysis of determinate physiological qualities – renders forensic practice as a 'triple system' of documentation, surveillance and automation which, via the introduction of the spectrograph, focused attention away from the voice as unique aspects of the individual, and towards a standardised framework for speech sounds which "began to provide composite templates for machine recognition" (Li and Mills 2019, 132).

Much has been written about the ways in which facial recognition technologies supervene upon phrenological and physiognomic assumptions (Stark and Hoey 2021). To grasp how machine learning adopts and augments this foundational strategy, we read the probabilistic impulse of vocal forensics through anthropologist Clyde Snow's method of osteobiography (the means by which he identifies an individual from their remains), taking it as a precursor to the predictive practice found in generative machine learning models, which seeks to identify new subjects on the grounds of their speculated remains. Amongst practitioners of counter-forensic, or investigative techniques Snow's work has been leveraged as a method to explore the ways in which the past can bear witness to the present (Keenan 2014), with objects, spaces, absences and gaps writing the biography of an incident. In many ways, the computational turn in vocal forensics seeks to cast the spectra of the voice signal in a similar light; a biographical source, bearing witness to the context within which the sample emerges and hypothesising about the speaker behind it. Going a step further, Speech2Face seeks not merely a forensic hypothesis of the signal, but a speculative reconstruction of its source, where the signal writes an autobiography of the uttering body.

Case Study

Multimodal Learning

The quality of a machine learning model typically depends on how effectively it can learn representations of the data upon which it is being trained. In the case of generative machine learning models, better representations of the salient features of the posterior data distribution are required to ensure that predicted outputs continue to fit the originally observed distribution (Alain et al. 2014). To understand the

claimed contribution of Speech2Face, it is helpful to understand the distinction between mono-modal and multi/cross-modal techniques in the context of representation learning. Modality here refers to the process of learning representations from a given type of data – audio, text, image, etc. Mono-modal tasks have been largely concerned with classification, aimed at annotating and tagging speech with estimations of demographic attributes such age (Zazo et al. 2018) and gender (Feld et al. 2010). Cross-modal approaches such as those deployed in Speech2Face operate across more than one modality of data in order to transform one into another. Speech2Face builds on previous cross-modal techniques for image retrieval, such as Kim et al. (2019), who propose a method for predicting which of two candidate portraits images a recording of speech is most likely to have originated from and Yan et al. (2016), who demonstrate a method for the generation of a portrait image based on provided visual attributes. Speech2Face's novel contribution to this set of methods was to join these two methods into a single pipeline for image generation.

While voice and image initially appear to exhibit differential modal structures (voice being a sequential, time-based audio signal whereas an image is a static, spatial arrangement of pixel values) machine learning techniques for the analysis of vocal signals builds upon the transformation of an audio signal into a spectrographic representation. This allows Speech2Face to combine both the association of faces and voices and the generation of novel facial portrait images. The task of generative facial reconstruction from these vocal signals relies upon developing an accurate mapping of these acoustic characteristics of the voice to various craniofacial parameters.

Critique

The authors note in their introduction that the project does not aim to reproduce a facsimile of the face of the speaking subject insofar as it is not concerned with identifying the speaker directly. Rather, the model aims at capturing the facial traits that can be positively associated with the vocal information found in speech. The authors go to some effort to clarify that Speech2Face should not be understood or used as a method of speaker identification in the forensic sense, emphasising its function as a method for revealing statistical correlations existing between features of speakers faces and their voices. In their statistical analysis the authors consider both demographic attributes including age, gender and ethnicity as well as similarity in landmark based craniofacial measurement such as "nose width", "upper lip height" and "nasal index". They compare the labels and values for these features by computing them from reconstructed input faces (rendered in profile) and the portraits produced by Speech2Face.

The data used to train the model consists of a collection of image and speech recording pairs. It does not feature demographic labels collected or otherwise inferred from the subjects. In the absence of a ground truth for demographic analysis, the authors turn to Face++, a commercially available face attribution classifier.

Researchers have demonstrated performance biases in these classifiers, Face++ being highlighted as one that performs demonstrably worse on darker skin (Buolamwini and Gebru 2018). Biases such

as these are brought into the correlations made here. The statistical evaluations of Speech2Face are not being made upon facts about the speaker and an estimate of the same attribute in the generated image, but rather, Speech2Face is evaluated on the basis of the extent to which the demographic labels inferred by a biased classifier – run on the generated images – correlate with demographic labels inferred by the same classifier, run on the input.

Similarly, to produce facial landmarks (measurements between facial features) for both "true" and generated faces both classes of image need to be in a "canonical position" (in profile). The generated faces are produced in this position, but the dataset images are not. In order to produce these measurements, the researchers turn to yet another machine learning model to generate an intermediate, reorientated representation of the face in the dataset. Again, Speech2Face is evaluated on the basis of how much the generated images it produces correlates with other synthetic images of faces, not real ones represented in the datasets.

In her discussion on "statistical renderings", Steyerl notes that the composites produced as part of the Racial Faces in the Wild Database, a set of "quasi-platonic" racial category portraits, "acquire the authority of an immediate manifestation or apparition [...t]hey skip mediation to gesture towards fake immanence" (2023). The portraits produced by Speech2Face function in a similar fashion. Features of the face that the authors argue are correlated with features of the voice are composed in the same image as those which are in no way correlated.

There is a circular logic at play here whereby, at points, the statistical correlations merely evaluate the generated images and at other points the generated images are simply vessels for the statistical correlations. The lack of clarity here allows the authors to present the portraits as the principal contribution but to fall back on the statistical correlations if the epistemic utility of the portraits is called into question. Further, we call into question whether any conclusions can be drawn from these correlations since they made on the basis of comparison with other hallucinated renderings and using biased facial classifiers in a form of recursive evaluation.

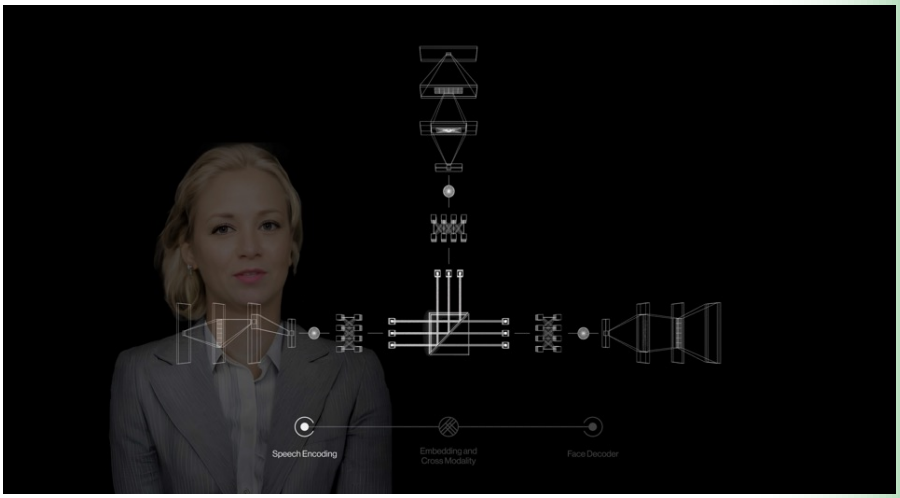

Fig. 1. "From model architecture to architectural model: speech2face deconstructed", Unit Test, *Not I*, 2023.

Film

The film is structured into three distinct chapters: a historical contextualisation, a deconstruction of the model, and a demonstration of an adversarial attack upon the model. The development of this narrative structure reflects a method of inquiry that leverages theoretical criticism, computational analysis and active intervention and evaluation. Following a historical framing that traces the development of vocal profiling technology and its relationship to the field of vocal forensics, a narrator introduces us to Speech2Face. Here, their disembodied voice is suddenly given a visual form, a human representation produced by running the audio of the narration through the model and animating the resulting face.

In this chapter the newly embodied presenter delivers a didactic deconstruction of the model's architecture by re-representing it as an architectural model. Speech2Face, like many machine learning models, must be understood as a socio-technical artefact – constructed in the context of a wider ecology of relations that inform its development beyond computational norms. Whilst this has been well established by studies focusing upon an analysis of the training dataset (Birhane et al. 2021), in order to open Speech2Face up to wider analysis about the representational capacity of the voice, we require a transformation in the scales of representation – from model architecture to architectural model. As Albert Smith suggests, the use of scale models within architecture allow practitioners to produce "an understandable surface (framework) upon which they can project and develop their measures of invisible things" (Smith 2007). Doing so not only makes apparent those elements of computational practice that are otherwise occluded by the functional remit of a model architecture, but also allows us to evaluate and interrogate Speech2Face with methods amenable to the obscured socio-cultural nature of its construction. Critical attention is paid to how the model draws the modalities of sight and sound into the same representational plane, encoding the assumption that faces which look the same should sound the same, and vice-versa.

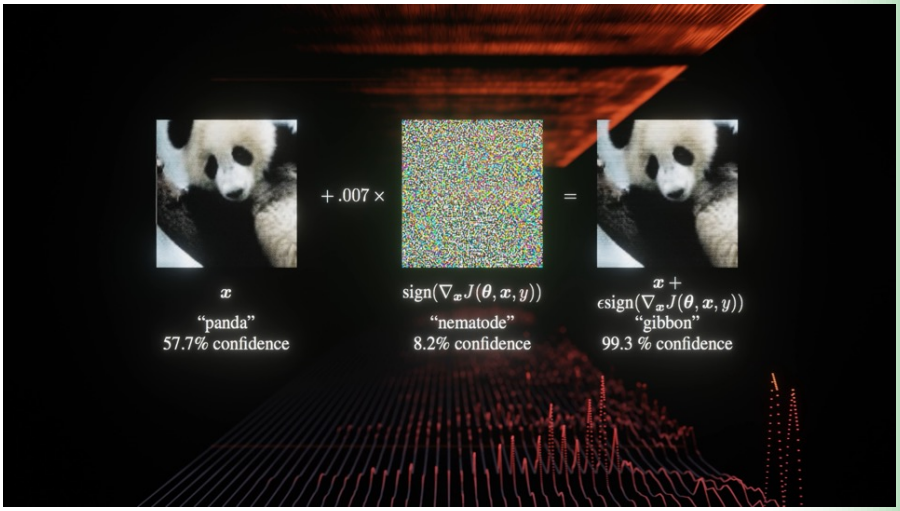

Fig. 2. "Adversarial attack", Unit Test, *Not I*, 2023.

The final chapter of the film demonstrates an adversarial attack upon Speech2Face. Using a bespoke machine learning method for al-

tering inputs to the model, subtle amounts of precisely generated noise are added. These perturbations are nearly imperceptible to the human ear yet cause the model to radically alter its output. Here, the narrator's voice is once again run through the Speech2Face pipeline, this time, as an adversarial example. At this point in the film another presenter appears within the frame, clearly different to the first, but also animated and lip synced to the voiceover mirroring the movements of the original. The two presenters then deliver the rest of the dialogue together.

Conclusion

> *If a single subject like me has voices how can there be a single 'the voice' to theorize? (Sterne 2019)*

Building upon the work of scholars across sound studies and vocal studies, Not I problematises an impulse in data science to apply statistical learning techniques to polymorphous features of human expression. Not I challenges the assumptions behind attempts to distil a one-to-one mapping of voice and identity, and in particular, reveals fundamental flaws in the attempt at speculative reconstruction demonstrated in Speech2Face.

We take a practice-based approach to evaluating the social assumptions at the heart of models such as Speech2Face, leveraging the affective qualities of moving image work to walk audiences through the use of contextual critique as a method for producing adversarial engagements with computational practice. The use of moving image becomes more than a mode of communication, more than a deconstruction of 'the fact that' these models are socio-culturally conditioned. Rather, we present this investigation as a form of experiential critique, forcing the viewer into an affective scenario that unsettles their existing heuristics used to infer speaker identity from vocal perception.

References

Alain, Guillaume, and Yoshua Bengio.
2014. "What Regularized Auto-Encoders Learn from the Data-Generating Distribution". *The Journal of Machine Learning Research* 15 (1): 3563–93.

Beckett, Samuel (dir.).
1972. "Not I". Forum Theatre, Lincoln Center, New York.

Birhane, Abeba, Vinay Uday Prabhu, and Emmanuel Kahembwe.
2021. "Multimodal Datasets: Misogyny, Pornography, and Malignant Stereotypes" 64 Not I. arXiv.http://arxiv.org/abs/2110.01963

Buolamwini, Joy, and Timnit Gebru.
2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In *Conference on Fairness, Accountability and Transparency*, 77– 91. PMLR.

Feld, Michael, Felix Burkhardt, and Christian Müller.
2010. "Automatic Speaker Age and Gender Recognition in the Car for Tailoring Dialog and Mobile Services". In *Eleventh Annual Conference of the International Speech Communication Association.*

Galton, Francis.
1879. "Composite Portraits, Made by Combining Those of Many Different Persons into a Single Resultant Figure." *The Journal of the Anthropological Institute of Great Britain and Ireland* 8: 132–44.

Keenan, Thomas.
"Getting the dead to tell me what happened: Justice, prosopopoeia, and forensic Afterlives". In *Forensis: The Architecture of Public Truth*. Edited by Eyal Weizman, Sternberg, 2014.

Kim, Changil, Hijung Valentina Shin, Tae-Hyun Oh, Alexandre Kaspar, Mohamed Elgharib, and Wojciech Matusik.
2018. "On Learning Associations of Faces and Voices". arXiv. https://doi.org/10.48550/arXiv.1805.05553

Li, Xiaochang, and Mara Mills.
2019. "Vocal Features: From Voice Identification to Speech Recognition by Machine". *Technology and Culture* 60 (2): S129–60.

McQuillan, Dan.
2018. "Data Science as Machinic Neoplatonism". *Philosophy & Technology* 31 (2): 253–72. https://doi.org/10.1007/s13347-017-0273-3

Oh, Tae-Hyun, Tali Dekel, Changil Kim, Inbar Mosseri, William T. Freeman, Michael Rubinstein, and Wojciech Matusik.
2019. "Speech2face: Learning the Face behind a Voice". In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7539–48.

Smith, Albert.
2007. *Architectural Model as Machine. Routledge.*

Stark, Luke, and Jesse Hoey.
2021. "The Ethics of Emotion in Artificial Intelligence Systems". In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 782–93.

Sterne, Jonathan.
2019. "Ballad of the Dork-o-Phone: Towards a Crip Vocal Technoscience". *Journal of Interdisciplinary Voice Studies* 4 (2): 179–89.

Steyerl, Hito.
2023. "Mean Images". *New Left Review*, no. 140/141 (April): 82–97.

Yan, Xinchen, Jimei Yang, Kihyuk Sohn, and Honglak Lee.
2016. "Attribute2image: Conditional Image Generation from Visual Attributes". In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, 776–91. Springer.

Zazo, Ruben, Phani Sankar Nidadavolu, Nanxin Chen, Joaquin Gonzalez-Rodriguez, and Najim Dehak.
2018. "Age Estimation in Short Speech Utterances Based on LSTM Recurrent Neural Networks". *IEEE Access* 6: 22524–30.