

**Speech-Derived Music:
Computational and Electromechanical
Methods in DIY Sound Art**

Fintan O'Hare

PhD

University of York

School of Arts and Creative Technologies

March 2026

Abstract

This practice research thesis develops an experimental artistic practice in which musical material is computationally derived from speech and realised through lo-fi, DIY electromechanical devices. The widespread use of speech as a compositional source points to its exceptional richness, including acoustic features as well as emotional, semantic, and syntactic qualities. While computational approaches to speech-derived music are well-established, their application within electromechanical sound-art remains underexplored. Through seven creative outputs, this research investigates how speech features can be transformed into computer-controlled acoustic sound, shaped by algorithmic processes and the material limitations of found objects.

Varying in terms of materials, interfaces and source speech-features, these works are structured in three stages reflecting the application of particular computational tools and methods: the algorithmic transcription and mapping of speech audio, the incorporation of Neural Audio Synthesis (NAS) as RAVE (Realtime Audio Variational autoEncoder) generated audio, and the use of text and language models to generate linguistic structures. Each investigates a unique speech feature, mapping method, and compositional application, reflecting on the technical and creative processes specific to its physical realisation. Improvisation is central to this process, from experimentation with objects and actuators, to electric motors and hardware, interfaces, performance, and composition.

The project finds that speech-derived features have applications to both compositional material and structure: as improvisational palettes, defining formal structures such as the gradual emergence of coherent text, and establishing generative musical systems that draw on underlying syntactic and semantic patterns. The act of translating high-dimensional computational features into low-resolution mechanical movement results in an approximate transcription of the source material. This imperfect rendering is developed as a generative method, in which algorithmic abstraction and the physical unpredictability of the devices create new musical outputs with traces of the original speech source.

The contributions include field reviews providing an overview of current practice; a set of compositional strategies for mapping speech and language to DIY electromechanical devices, methods for the performative shaping of algorithmic output, and techniques for using syntax and semantic data as a non-standard control signal for musical structure; publicly available code for each method and device; and RAVE audio and LSTM text models trained throughout the research. These methods provide practical ways of working with speech in experimental artistic practice, offering future directions for speech-derived music, sound-art, and DIY electromechanical performance.

Funding

This work was supported by the Arts & Humanities Research Council (grant number AH/R012733/1) through the White Rose College of the Arts & Humanities (WRoCAH).

Acknowledgements

Dr. Federico Reuben for supervision, encouragement and support.

Jo Cross, Dina Hashem, Clare Meadley and all at the White Rose College of Arts and Humanities (WRoCAH) for their support, training, funding, and guidance.

All staff at the School of Arts and Creative Technologies at the University of York for opportunities to present and share my work.

Gaia Blandina for a rewarding improvisatory collaboration, and Danny Saleeb, Jack McNeill, Bethany Nichol, Alex Mackay, and James Wood for their help in playing and recording.

Wonkystuff, Navigators Art and the Wild Research Symposium for a place to share my work. And Nick, for supplying me with new objects to play with.

Lewis and Emm, for encouraging me to do this in the first place.

Leila, for love and good food.

Declaration

I declare that this thesis is a presentation of original work and I am the sole author, except for the collaborative improvisational works clearly labelled in Chapter 4. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

Chapter 2 is based on the following journal paper.

Fintan O'Hare. Computational Methods in Speech-Derived Music: Field Review.
Leonardo, 58(4):409-417, 2025.

4.3 was presented at the *Wild Research Symposium* (Journal *Media Practice and Education* and The Revelator Wall of Death). 14 Sep 2024.

5.3 was presented at *Algorithmic Audiovisualities: Music and Media in the Age of AI* (CysMus / CESEM). *Radio User Guide: Textual and Mechanical Limitations in Audiovisual Composition*. 13 Nov 2025.

Contents

List of Figures	x
List of Portfolio Examples	xii
List of Code Examples	xiii
List of Model Examples	xiii
Structure	xiv
1 Introduction	1
1.1 Speech-Derived Music	1
1.2 Motivation	2
1.3 Speech to Mechanical Devices	3
1.4 Speech Features	4
1.4.1 Acoustic Features	5
1.4.2 Linguistic Features	5
1.5 Audio and Text Models	6
1.5.1 RAVE Models	6
1.5.2 Text Models	7
1.6 Research Questions and Objectives	8
1.7 Methodology	9
1.7.1 Field Review	10
1.7.2 Speech Sources	10
1.7.3 Feature Selection and Mapping	10
1.7.4 Devices	11
1.7.5 Improvisation	12
1.8 Contributions	12
1.9 Overview of Contents	13
1.10 Conclusion	18
2 Computational Methods in Speech-Derived Music	19
2.1 Introduction	19
2.2 Methodology	21
2.3 Results	22
2.4 Computational Methods: Selected Works	25
2.4.1 Sample Playback	25
2.4.2 Electronic Manipulation	26
2.4.3 Mapping	28
2.4.4 Speech Synthesis	30
2.5 Discussion	31

2.5.1	Computational Methods	31
2.5.2	Sound Sources	32
2.5.3	Artistic Intent	33
2.5.4	Instrumentation	34
2.5.5	Compositional Challenges	35
2.5.6	Trends	35
2.5.7	Previous Research	36
2.5.8	Limitations	37
2.6	Conclusion	37
3	Machine Learning in Electromechanical Sound Art	39
3.1	Introduction	39
3.2	Electromechanical Music	41
3.3	Methodology	42
3.4	Results	42
3.4.1	Audio Classification	43
3.4.2	MIDI Generation	44
3.4.3	Live Parameter Mapping	46
3.4.4	System Modelling	47
3.4.5	Analogue Neural Networks	47
3.4.6	Speech Transcription	48
3.4.7	Text and Language Generation	48
3.5	Discussion	49
3.5.1	Current and Emerging Applications	50
3.5.2	Material Constraints	53
3.5.3	Interaction	54
3.6	Conclusion	56
4	Algorithmic Transcription in Practice	58
4.1	Introduction	58
4.2	MotorCase I	60
4.2.1	Overview	60
4.2.2	Speech Source	61
4.2.3	Device	62
4.2.4	Transcription Method	63
4.2.5	Improvisation	65
4.2.6	Composition	67
4.2.7	Evaluation	69
4.2.8	Summary	73
4.3	Motorcase II	74
4.3.1	Preamble	74
4.3.2	Motorcase II: Overview	75
4.3.3	Speech Source	76
4.3.4	Device	77
4.3.5	Transcription Method	78
4.3.6	Improvisation	79
4.3.7	Composition	81
4.3.8	Evaluation	82
4.3.9	Summary	86

4.4	Discussion	86
4.4.1	Methods of Abstraction	87
4.4.2	Material Realisation	87
4.4.3	Interaction and Performance	90
4.4.4	Conclusion	91
5	Audio Models in Practice	92
5.1	Introduction	92
5.2	Smalltalk	94
5.2.1	Overview	94
5.2.2	Dataset and Training	95
5.2.3	Synthesis	96
5.2.4	Organisation	97
5.2.5	Device	98
5.2.6	Composition	99
5.2.7	Evaluation	100
5.2.8	Summary	103
5.3	Radio User Guide	104
5.3.1	Overview	104
5.3.2	Dataset	105
5.3.3	Training and Synthesis	106
5.3.4	Device	106
5.3.5	Organisation	107
5.3.6	Composition	110
5.3.7	Evaluation	112
5.3.8	Summary	115
5.4	Discussion	115
5.4.1	Generating Instability	115
5.4.2	Methods of Organisation	116
5.4.3	Material Integration	117
5.4.4	Conclusion	118
6	Text Models in Practice	119
6.1	Introduction	119
6.2	Porcelain Music Box	121
6.2.1	Overview	121
6.2.2	Linguistic Structure	121
6.2.3	Mapping	123
6.2.4	Device	124
6.2.5	Interface	125
6.2.6	Evaluation	126
6.2.7	Summary	128
6.3	Seven Studies for Electric Motors	130
6.3.1	Overview	130
6.3.2	Linguistic Structure	131
6.3.3	Mapping	132
6.3.4	Device	133
6.3.5	Improvisation	135
6.3.6	Evaluation	137

6.3.7	Summary	139
6.4	Word games	140
6.4.1	Overview	140
6.4.2	Linguistic Structure	140
6.4.3	Mapping	141
6.4.4	Device	144
6.4.5	Composition	146
6.4.6	Evaluation	148
6.4.7	Summary	150
6.5	Discussion	151
6.5.1	Mapping Linguistic Structure	151
6.5.2	Material Constraints	152
6.5.3	Text Models and Agency	153
6.5.4	Conclusion	154
7	Conclusions	156
7.1	Introduction	156
7.2	Summary	157
7.3	Practice Overview	159
7.4	Themes	162
7.4.1	Approximation	162
7.4.2	Material Constraints	162
7.4.3	Interaction	164
7.4.4	Transparency	165
7.5	Contributions	165
7.6	Limitations	166
7.7	Future Work	167
	Appendices	169
A	Devices	169
A.1	Motorcase I	169
A.2	Motorcase II	170
A.3	Smalltalk	170
A.4	Radio User Guide	171
A.5	Porcelain Music Box	171
A.6	Seven Studies	172
A.7	Word Games	172
	References	173

List of Figures

4.1	<i>Motorcase I</i> device, showing the found objects and actuators	62
4.2	<i>Motorcase I</i> hardware	62
4.3	Visualisation of the text-based rhythmic segmentation	64
4.4	Individual sample controls	66
4.5	Global controls	66
4.6	Plot of speech onsets within the first 10 seconds of each transcription, showing variation in density and spacing.	70
4.7	Plot of transcription data within the first 10 seconds of the ‘Ted talk’ audio, showing the levels of the text-based segmentation.	70
4.8	Spectrogram of short looping sequence on the device, showing variation due to variability in the object-actuator configurations. Loop marked with dashed orange lines.	71
4.9	<i>Motorcase I</i> transcription showing differences between speech, motor data, and motor realisation in the letters segmentation	72
4.10	<i>MotorCase II</i> , open and closed.	77
4.11	<i>MotorCase II</i> , hardware	77
4.12	<i>Motorcase II</i> MIDI transcription (orange) overlaid onto a spectrogram of the original audio (grey), showing the relationship between the transcribed data and speech source.	83
4.13	Heatmap showing changes to average onsets per second across the VAD space of the MIDI transcriptions	84
4.14	Heatmap showing changes to note duration range across the VAD space of the MIDI transcriptions	84
4.15	Normalised LTAS of <i>Motorcase I</i> and <i>II</i> . <i>Motorcase I</i> (orange) exhibits a broader distribution resulting from its mixed material palette, while <i>Motorcase II</i> (green) shows more distinct peaks with its all-metal construction.	88
4.16	Mechanism 1, a rotating wheel which produces a rising-falling sound (green). . .	89
4.17	Mechanism 2, a rotating arm striking two objects to produce two distinct attacks (blue).	89
4.18	The realisation, in which the sound of each mechanism is superimposed on the form of the transcribed data. The transcription here is represented simply as a sequence of two events, mapped directly to each mechanism.	89
5.1	<i>Smalltalk</i> drawing interface	98
5.2	<i>Smalltalk</i> devices	99
5.3	<i>Smalltalk</i> design	99
5.4	<i>Smalltalk</i> contour matching, in which simple shapes broadly correspond to retrieved audio in the <i>Portfolio</i>	101
5.5	<i>Smalltalk</i> manually-transcribed MIDI (top) transcription from audio (bottom) .	102
5.6	One of three devices	107

5.7	<i>Radio User Guide</i> design	107
5.8	<i>Radio User Guide</i> score key	111
6.1	<i>Porcelain Music Box</i> program structure	123
6.2	Patterns of Continuation and Disruption	124
6.3	<i>Porcelain Music Box</i> design	125
6.4	<i>Porcelain Music Box</i> device	126
6.5	MRDA corpus and LSTM distribution similarities	127
6.6	Rhythmic onsets from the video recording, showing patterns of dense and sparse material	127
6.7	<i>Seven Studies</i> audio from speech recordings analysed for pitch and amplitude, translated into a 4x4 grid where point size represents motor speed.	132
6.8	<i>Seven Studies</i> example grids	133
6.9	<i>Seven Studies</i> design	134
6.10	<i>Seven Studies</i> objects and actuators	135
6.11	<i>Seven Studies</i> controller settings	137
6.12	Example of <i>Word games</i> 7-dimensional embeddings, visualised using coloured squares to show similarities between related words	142
6.13	<i>Word games</i> curve interpolation for a single embedding dimension; the initial set of curves were chosen manually.	142
6.14	<i>Word games</i> design	144
6.15	<i>Word games</i> device with camera	145
6.16	Relative simplicity of <i>Word games</i> objects and actuators to <i>Seven Studies</i>	149
A.1	Motorcase I	169
A.2	Motorcase II	170
A.3	Smalltalk	170
A.4	Radio User Guide	171
A.5	Porcelain Music Box	171
A.6	Seven Studies	172
A.7	Word Games	172

List of Tables

2.1	Computational methods, categories and works.	22
2.2	Computational methods, number of works in each category.	31
2.3	Number of works by sound source.	32
2.4	Number of works by shared theme.	33
2.5	Number of works by instrumentation.	34
3.1	Overview of surveyed works	51
4.1	Overview of Project Elements: <i>Motorcase I</i> and <i>Motorcase II</i>	59
4.2	Abstraction and Effect in <i>Motorcase I</i> and <i>II</i>	87
4.3	Performance Controls for <i>Motorcase I</i> and <i>II</i>	90
5.1	Overview of Project Elements: <i>Smalltalk</i> and <i>Radio User Guide</i>	93
5.2	Example LSTM text outputs.	108
5.3	<i>Radio User Guide</i> word association examples.	109
5.4	Organisational Methods in <i>Smalltalk</i> and <i>Radio User Guide</i>	116
6.1	Overview of Project Elements: Text Models in Practice	120
6.2	Comparison of Linguistic Mapping and Sequence Generation	152
7.1	Computational methods and works	160
7.2	Works by speech source.	161
7.3	Summary of Devices.	161

List of Portfolio Examples

4.1	<i>Motorcase I</i> : Objects and Actuators	63
4.2	<i>Motorcase I</i> : Improvisation in Practice	66
4.3	<i>Motorcase I</i> : Video Composition	68
4.4	<i>Motorcase I</i> : Transcription Examples	71
4.5	<i>Motorboard</i> : Video	74
4.6	<i>Motorcase II</i> : Improvisation Examples	80
4.7	<i>Motorcase II</i> : Composition Examples	82
4.8	<i>Motorcase II</i> : Transcription Example	83
5.1	<i>Smalltalk</i> : Video	100
5.2	<i>Smalltalk</i> : Contour Matching	101

5.3 <i>Radio User Guide</i> : Word Association	109
5.4 <i>Radio User Guide</i> : Collage	110
5.5 <i>Radio User Guide</i> : Feedback Loops	111
5.6 <i>Radio User Guide</i> : Performance	111
5.7 <i>Radio User Guide</i> : Dataset to Synthesis	113
6.1 <i>Porcelain Music Box</i> : Video	126
6.2 <i>Seven Studies</i> : Video	137
6.3 <i>Word games</i> : Video	147

List of Code Examples

4.1 <i>Motorcase I</i> : Transcription Detail	64
4.2 <i>Motorcase I</i> : Interface Detail	65
4.3 <i>Motorcase I</i> : Video Detail	67
4.4 <i>Motorcase II</i> : Dataset and Navigation	76
4.5 <i>Motorcase II</i> : Transcription Preprocessing	78
4.6 <i>Motorcase II</i> : Interface Detail	79
5.1 <i>Smalltalk</i> : Contours	97
5.2 <i>Smalltalk</i> : Communication	98
6.1 <i>Porcelain Music Box</i> : Script	124
6.2 <i>Seven Studies</i> : Scripts	136
6.3 <i>Word games</i> : Mapping	143

List of Model Examples

4.1 <i>Motorcase II</i> : RAVE models	81
5.1 <i>Smalltalk</i> : Models	96
5.2 <i>Smalltalk</i> : Model Audio	101
5.3 <i>Radio User Guide</i> : Dataset	105
5.4 <i>Radio User Guide</i> : Models	106
5.5 <i>Radio User Guide</i> : Synthesis to Organisation	113
6.1 <i>Porcelain Music Box</i> : Model	123

Structure

- *Portfolio* contains recordings and scores of the creative works.
- *Code* contains code for the devices and compositional methods.
- *Models* contains RAVE and text models trained for this research.

Note on Examples

In this commentary I use shaded boxes (like this one) to point to examples from the submitted *Portfolio*, *Code* and *Models* folders. Footnotes^a indicate the file path in the repository, icons provide online links for convenience.

► Recordings: <https://doi.org/10.6084/m9.figshare.31324327>

≡ Scores: <https://doi.org/10.6084/m9.figshare.31324327>

</> Code: https://github.com/fifoh/-Speech_Derived_Music_Code

📄 Models: <https://huggingface.co/fifoh/speech-derived-music-models/tree/main>

^a. for example, *Portfolio/Chapter 4 - Motorcase I/...*

Chapter 1

Introduction

1.1 Speech-Derived Music

What is speech-derived music? The term is often associated with the intuitive transcription and instrumental imitation of speech found in the music of Steve Reich¹ or Scott Johnson.² But speech-melody transcription is only one approach. Composers and sound artists have derived musical material from speech in multiple ways, drawing on acoustic features, emotional, semantic and syntactic qualities among others.³ This broad definition includes Alessandro Bosetti's *Mask Mirror*,⁴ for example, which reorganises sampled speech in random sequences; Benjamin Bacon's *PROBE II, Subaudition*⁵ which transcribes and transmits an audience's speech to a zoetrope projection; Mauro Lanza's *The Kempelen Machine*⁶ which uses computer-controlled accordion reeds.

Even in this small group of examples, connections between mechanical instruments and

1. Eric Prieto, 'Speech Melody and the Evolution of the Minimalist Aesthetic in Steve Reich's The Cave', *Circuit* 12, no. 2 (2002): 21–44, <https://doi.org/10.7202/902250ar>.

2. Jonathan Maltman and Lindsay Vickery, 'The Foundations of Speech-Melody Composition And The Dawn Of The Digital Music Era', in *Proceedings of the Australasian Computer Music Conference* (2018), 30–37.

3. Cathy Lane, 'Voices from the Past: Compositional approaches to using recorded speech', *Organised Sound* 11, no. 1 (2006): 3–11, <https://doi.org/10.1017/s1355771806000021>; Lauren Redhead, 'Language as a Material Process', *Contemporary Music Review* 39, no. 3 (2020): 359–372, <https://doi.org/10.1080/07494467.2020.1821527>; Daniel Buner Formo, 'The Orchestra of Speech—a speech-based instrument system', in *Proceedings of the International Conference on New Interfaces for Musical Expression* (2018), <https://doi.org/10.5281/zenodo.1302607>.

4. Alessandro Bosetti, *Mask Mirror*, 2008, accessed 25 February 2026, <https://melgun.net/travaux/maskmirror/>.

5. Benjamin Bacon, *PROBE II, Subaudition*, 2021, accessed 25 February 2026, <https://www.benjaminbacon.studio/probe-1>.

6. Mauro Lanza, *The Kempelen Machine*, 2015, accessed 25 February 2026, <https://brahms4.ircam.fr/fr/works/work/41088/>.

speech begin to emerge. Early examples—like Lanza’s inspiration, von Kempelen’s eighteenth century speaking machine⁷—were designed to reproduce speech, but the uncanny speech yet non-speech sound qualities of these machines can also be found in recent composition and sound art. From Peter Ablinger’s abstraction and ‘sudden shift into recognition’ in the computer-controlled piano of *Quadraturen III*⁸ to Jesus Canuto’s stepper-motor rendering of speech in *Talking Motors*,⁹ the boundary between speech and non-speech is often deliberately blurred.

Recent computational approaches to speech, including in feature extraction¹⁰ and synthesis¹¹ suggest opportunities for developing novel approaches to speech-derived music. At the same time, artists and composers have been broadening their available sound qualities through DIY electromechanical instruments,¹² using electric motors, actuators and microcontrollers. By combining computational speech analysis with lo-fi motor-driven instruments, this work contributes to the field of practice research in sound art and contemporary music, developing and reflecting on an experimental artistic practice of deriving musical material from speech, and its realisation through motors and found objects.

This commentary is structured in two parts: field review, followed by documentation and reflection on the practice itself. The *Portfolio* presents the creative outputs of this research, with this commentary providing context for the works and highlighting specific elements. *Code* and *Models* contain supplementary material. For convenience, online links to media ► code </> models ◻ and scores ≡ are given.

1.2 Motivation

My central motivation in this work is of play—with speech, language, and DIY machines. This follows a longstanding personal interest in the musical qualities of speech, transcribing, notating,

7. Slavo Ondrejovic, ‘Wolfgang von Kempelen and His Speaking Machine’, *Human Affairs* 2, no. 2 (1992): 161–172, <https://doi.org/10.1515/humaff-1992-020208>.

8. Peter Ablinger, *Quadraturen*, 1995 -, accessed 25 February 2026, <https://ablinger.mur.at/docu11.html>.

9. Jesus Canuto, *Talking Motors*, 2023, accessed 25 February 2026, <https://www.jesuscanto.com/talking-motors>.

10. Douglas O’Shaughnessy, ‘Review of analysis methods for speech applications’, *Speech Communication* 151 (2023): 64–75, <https://doi.org/10.1016/j.specom.2023.05.008>.

11. Xu Tan et al., ‘A Survey on Neural Speech Synthesis’, *arXiv*, 2011, 2106.15561, <https://doi.org/10.48550/arXiv.2106.15561>, eprint: 2106.15561.

12. Jon Pigott, ‘Across Fields: Sound, art and technology from an electromechanical perspective’, *Organised Sound* 22, no. 2 (2017): 276–285, <https://doi.org/10.1017/s1355771817000188>; Asbjørn Blokkum Flø, ‘Materiality in sound art’, *Organised Sound* 23, no. 3 (2018): 225–234, <https://doi.org/10.1017/s1355771818000134>; Mo H Zareei, ‘Audiovisual materialism’, *Organised Sound* 25, no. 3 (2020): 362–371, <https://doi.org/10.1017/s1355771820000321>.

and incorporating spoken words and syllables into compositions. This is partly aesthetic preference, as I find many of the sound qualities of speech are also recurring compositional concerns: noise-based sounds, angular rhythmic structures, timbral variation and repetition. However, this is also driven by an interest in the space between speech and music, explored through feature extraction and translation to mechanical devices.

I see this as analogous to the visual practice of asemic writing, which Peter Schwenger describes as ‘writing that locates its essence in the rhythmic and gestural relations of marking’.¹³ With sense-making stripped from speech, focus can be directed to sonic, gestural, emotional or structural features, as well as qualities which emerge from the mapping process and mechanical devices themselves. While these processes might distance consequent material from its source, the features of speech—extracted and transformed—continue to influence the resulting work.

1.3 Speech to Mechanical Devices

Early mechanical attempts to recreate the speaking voice may not seem convincingly human,¹⁴ but are interesting aesthetically because of this reason: they present an imperfect representation in which some sonic features of speech remain, while others emerge from the limitations of the mechanism.¹⁵ Incorporating electromechanical elements—motors, solenoids—into speech-derived music has grown in recent years, enabled by increased accessibility to hardware and software, including open-source tools and educational resources.¹⁶ As with the early examples, an aesthetic appeal is in the difference between the speech source and its mechanical realisation. Notable works include Ablinger’s *Quadraturen III*,¹⁷ which maps speech to computer-controlled player piano, Formo’s *The Orchestra of Speech*,¹⁸ which resynthesises speech using transducers

13. Peter Schwenger, *Asemic: The art of writing* (University of Minnesota Press, 2019), 66.

14. Ondrejovic, ‘Wolfgang von Kempelen and His Speaking Machine’; Riley McGuire, ‘Writing Novels, Simulating Voices: Euphonia, Trilby, and the Technological Sounding of Identity’, *Victorian Literature and Culture* 49, no. 2 (2021): 325–360, <https://doi.org/10.1017/s1060150319000251>; M Helmholtz, ‘XII. On vowel sounds’, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 19, no. 125 (1860): 81–88, <https://doi.org/10.1080/14786446008642829>.

15. Gordon J Ramsay, ‘Mechanical speech synthesis in early talking automata’, *Acoustics Today* 15, no. 2 (2019): 11–19, <https://doi.org/10.1121/at.2019.15.2.11>.

16. Flø, ‘Materiality in sound art’; Zareei, ‘Audiovisual materialism’.

17. Peter Ablinger, *Quadraturen III*, 2004, accessed 25 February 2026, <https://ablinger.mur.at/docu11.html#qu3>.

18. Formo, ‘The Orchestra of Speech—a speech-based instrument system’.

mounted on instruments, and Villicaña-Shaw's *Robot Whispers*,¹⁹ which maps speech frequency content to instrument actuators. DIY methods using recycled materials can be found in Andrea Valle's *Rumentarium*, in which Valle notes that the limitations of the devices should be seen as a creative affordance, forcing the artist to explore the object's intrinsic constraints.²⁰

These mechanical instruments introduce physical and material dimensions that influence both the sound and the conceptual approach, and I would suggest that considerations of physical constraints, actuation mechanisms, and the resonant properties of materials often provoke unique approaches to the representation of speech. This aligns with broader concerns present in data mapping across music and sound art, where choices about which features in the data to emphasise or discard are critical to the creative process.²¹

1.4 Speech Features

Composers and sound artists draw on a range of external sources beyond speech, from radio waves in Quadrature's *Noise Signal Silence*²² to tidal patterns in Eacott's *Floodtide*.²³ So why is speech the focus of this practice research? It is partly a matter of its richness as a source: speech offers a range of possible features to choose between, from signal feature-extraction to emotional qualities and linguistic annotation. It seems that few, if any, other sources offer comparable tools for analysis and feature extraction, and breadth of research into synthesis and generative techniques.²⁴ These possibilities allow for several distinct compositional approaches, the first of which concerns features drawn from the speech-signal.

19. Nathan Villicaña-Shaw, Spencer Salazar and Ajay Kapur, 'Mechatronic Performance in Computer Music Compositions', in *Proceedings of the International Conference on New Interfaces for Musical Expression* (2018), 413–418, <https://doi.org/10.5281/zenodo.1302647>.

20. Andrea Valle, 'Making acoustic computer music: The Rumentarium project', *Organised Sound* 18, no. 3 (2013): 242–254, <https://doi.org/10.1017/s1355771813000216>.

21. Paul Doornbusch, 'Composers' views on mapping in algorithmic composition', *Organised sound* 7, no. 2 (2002): 145–156, <https://doi.org/10.1017/s1355771802002066>; Lev Manovich, *Data Visualization as New Abstraction and Anti-Sublime*, 2002, accessed 25 February 2026, https://manovich.net/content/04-projects/041-data-visualisation-as-new-abstraction-and-anti-sublime/37_article_2002.pdf.

22. Quadrature, *Noise Signal Silence*, 2019, accessed 25 February 2026, <https://quadrature.co/work/noise-signal-silence/>.

23. John Eacott, 'Flood Tide: sonification as musical performance—an audience perspective', *AI & society* 27 (2012): 189–195, <https://doi.org/10.1007/s00146-011-0338-2>.

24. Fabio Cifariello Ciardi, 'Strategies and tools for the sonification of prosodic data: A composer's perspective', in *Proceedings of the 26th International Conference on Auditory Display (ICAD 2021)* (2021), 247–252, <https://doi.org/10.21785/icad2021.041>; Rébecca Kleinberger et al., 'Voice at NIME: a Taxonomy of New Interfaces for Vocal Musical Expression', in *Proceedings of the International Conference on New Interfaces for Musical Expression* (2022), <https://doi.org/10.21428/92fb44.4308fb94>.

1.4.1 Acoustic Features

Speech-signals carry timbral, rhythmic, and dynamic characteristics that can be extracted and repurposed musically. Low-level acoustic features consist of statistical descriptors derived from the audio waveform,²⁵ such as the spectral centroid, flatness, flux, or amplitude envelope. Higher-level features are closer to human perception,²⁶ and include rhythmic onsets, MFCC's, or fundamental frequency tracking. Each offers a possible starting point to develop methods for the feature extraction and mapping of speech to electromechanical devices.

Elements of these signals contribute to broader perceptual features:²⁷ emotional speech, for example, is characterised by exaggerated rhythmic and dynamic fluctuations, and the emotions of fear and happiness can be distinguished through a comparison of their energy contour.²⁸ Research has suggested shared acoustic features across speech and music emotion,²⁹ and even when stripped of semantic content, these features can retain perceptual associations with speech-like qualities.³⁰ To me, this suggests an opportunity for deriving musical material specifically from emotional speech, with some expectation of shared sonic cues between the speech source and musical realisation.

1.4.2 Linguistic Features

Parallel to the acoustic features of speech are linguistic features. These are often documented through linguistic annotation, with a variety of computational tools and annotated datasets from constituency and phrase dependency grammars to dialogue acts.³¹ For example, the Meeting Recorder Dialog Act (MRDA) corpus³² consists of multi-party conversations of real-world meet-

25. Garima Sharma, Kartikeyan Umamathy and Sridhar Krishnan, 'Trends in audio signal feature extraction methods', *Applied Acoustics* 158 (2020): 107020, <https://doi.org/10.1016/j.apacoust.2019.107020>.

26. Sharma, Umamathy and Krishnan.

27. Bernard Gold, Nelson Morgan and Daniel Ellis, *Speech and audio signal processing: processing and perception of speech and music* (Wiley, 2011), 261.

28. K Sreenivasa Rao, Shashidhar G Koolagudi and Ramu Reddy Vempada, 'Emotion recognition from speech using global and local prosodic features', *International journal of speech technology* 16 (2013): 143–160, <https://doi.org/10.1007/s10772-012-9172-2>.

29. Eduardo Coutinho and Nicola Dibben, 'Psychoacoustic cues to emotion in speech prosody and music', *Cognition & emotion* 27, no. 4 (2013): 658–684, <https://doi.org/10.1080/02699931.2012.732559>.

30. Roger Dean, 'NoiseSpeech, a Noise of Living Bodies: Towards Attali's' Composition'', *Journal of New Media and Culture* 3, no. 2 (2005): 1–12, accessed 25 February 2026, <https://ibiblio.org/nmediac/winter2004/NoiseSp.htm>.

31. Stefan Th. Gries and Andrea L. Berez, 'Linguistic Annotation in/for Corpus Linguistics', in *Handbook of Linguistic Annotation*, ed. Nancy Ide and James Pustejovsky (Springer, 2017), 379–409, https://doi.org/10.1007/978-94-024-0881-2_15.

32. Elizabeth Shriberg et al., 'The ICSI meeting recorder dialog act (MRDA) corpus', in *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004* (2004), 97–100, <https://doi.org/10.21236/ada460980>.

ings labelled according to speaker identity and dialogue act, and the Penn Treebank³³ contains part-of-speech (POS) tagged text from articles to telephone conversations. From a compositional perspective these can provide a way of organising sound, using borrowed linguistic structures.³⁴

1.5 Audio and Text Models

Acoustic features and linguistic annotation both concern pre-existing speech sources, often involving audio feature-extraction or annotated datasets as discussed. But advances in audio and text generation present an alternative approach: using generated speech audio and text as the basis for new creative work.

1.5.1 RAVE Models

Neural Audio Synthesis (NAS) represents an advance from previous methods of formant synthesis and waveform concatenation.³⁵ In NAS, deep neural networks are trained on raw audio data to generate audio with the characteristics of the dataset. RAVE³⁶ (Real-time Audio Variational autoEncoder) was chosen for this project as a state-of-the-art system capable of timbral transfer between non-speech and speech, as well as enabling improvisatory exploration of a model's latent space³⁷—a multi-dimensional statistical representation of learned audio features.

Nicola Privato situates his use of RAVE models within a practice of *AI hauntography*, framing it as a continuation of the sonic hauntology movement through the deliberate magnification of the technological uncanny:

‘... we reveal the statistical flattening of the source material and the processes

33. Mitch Marcus, Beatrice Santorini and Mary Ann Marcinkiewicz, ‘Building a large annotated corpus of English: The Penn Treebank’, *Computational linguistics* 19, no. 2 (1993): 313–330, <https://doi.org/10.21236/ada273556>.

34. Jiin Ko, Youngjun Choi and Jinjoon Lee, ‘Harmonic Words, Narrative Chords: Textual Sonification Using Part of Speech’, in *Proceedings of the International Conference on New Interfaces for Musical Expression* (2024), 28–32, <https://doi.org/10.5281/zenodo.13904771>; François-Bernard Mâche, *Korwar*, 1972, accessed 25 February 2026, <https://brahms.ircam.fr/work/korwar>; Clarence Barlow, ‘On Music Derived from Language’, in *Proceedings of the 3rd International Symposium on Systems Research in the Arts and Humanities, Baden-Baden, Germany* (2009), accessed 25 February 2026, <https://clarlow.org/wp-content/uploads/2025/06/On-Music-Derived-from-Language.pdf>.

35. Yishuang Ning et al., ‘A review of deep learning based speech synthesis’, *Applied Sciences* 9, no. 19 (2019): 4050, <https://doi.org/10.3390/app9194050>.

36. Antoine Caillon and Philippe Esling, ‘RAVE: A variational autoencoder for fast and high-quality neural audio synthesis’, *CoRR*, 2021, <https://doi.org/10.48550/arXiv.2111.05011>.

37. Franziska Schroeder and Federico Reuben, ‘Breathing Life into AI: Querying AI through Tactile Imprints Using Breath and Sound’, in *Artificial Intelligence in the Cultural and Creative Sectors* (Routledge, 2025), 37–54, <https://doi.org/10.4324/9781003509738-5>.

activated by its algorithmic manipulation.’³⁸

This amplification of emergent qualities which distinguish generated audio also has much in common with the use of mechanical devices: the results are as much shaped by the process and the medium of representation as they are by the source.

1.5.2 Text Models

Alongside speech audio synthesis, text generation has undergone a rapid development.³⁹ Early approaches relied on Markov chains,⁴⁰ an approach limited by a narrow context of preceding words. Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) networks,⁴¹ addressed this through an internal memory state, offering a method for long-range sequence generation. Recent advances have shifted toward Large Language Models (LLMs) based on the Transformer architecture.⁴² While this improves coherence across long passages, the immense computational requirements of LLMs make them difficult to embed in DIY hardware. Small Language Models⁴³ (SLMs) offer a solution by retaining generative quality within a smaller parameter footprint—often through distillation, where a compact model is trained to reproduce the output of a larger LLM—making them more suitable for local deployment. For details of each model architecture see Jurafsky and Martin.⁴⁴

As text generation methods have developed, composers and artists have repurposed these technologies in a musical context. For example, Muenz’ 2004 *deChiffraGE*⁴⁵ stochastically shuffles letters and syllables to affect the gradual emergence of coherent text; LSTM networks

38. Nicola Privato and Thor Magnusson, ‘Querying the Ghost: AI Hauntography in NIME’, in *Proceedings of the International Conference on New Interfaces for Musical Expression* (2024), 432–438, <https://doi.org/10.5281/zenodo.13904901>.

39. Jonas Becker et al., ‘Text generation: A systematic literature review of tasks, evaluation, and challenges’, *arXiv preprint arXiv:2405.15604*, 2024, <https://doi.org/10.48550/arXiv.2405.15604>.

40. Claude Elwood Shannon, ‘A mathematical theory of communication’, *The Bell system technical journal* 27, no. 3 (1948): 379–423, <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>.

41. Alex Graves, ‘Generating sequences with recurrent neural networks’, *arXiv preprint arXiv:1308.0850*, 2013, <https://doi.org/10.48550/arXiv.1308.0850>.

42. Junyi Li et al., ‘Pre-trained language models for text generation: A survey’, *ACM Computing Surveys* 56, no. 9 (2024): 1–39, <https://doi.org/10.1145/3649449>.

43. Fali Wang et al., ‘A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness’, *ACM Transactions on Intelligent Systems and Technology* 16, no. 6 (2025): 1–87, <https://doi.org/10.1145/3768165>.

44. Daniel Jurafsky and James H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*, 3rd (2026), accessed 25 February 2026, <https://web.stanford.edu/~jurafsky/slp3/>.

45. Harald Muenz, *deChiffraGE*, 2004, accessed 25 February 2026, <https://www.musicwedliketohear.com/pdf/29032006.pdf>.

have been used extensively for melody and chord generation,⁴⁶ MIDI generation,⁴⁷ and text for art installations;⁴⁸ LLMs were used to generate text scores for Jennifer Walshe’s 2021 *Text Score Dataset*.⁴⁹ In practice, LSTM models offer training from custom datasets—which can be much smaller than those required for LLMs—from semantic text to linguistic patterns. SLMs on the other hand make the generation of higher-quality syntactically coherent text possible on embedded devices, suggesting possibilities for the real-time synthesis of novel and coherent text within an instrument itself.

1.6 Research Questions and Objectives

This practice research investigates the translation of speech to electromechanical sound through three computational approaches, and result in the following research questions:

- R1. Which features can be computationally extracted from the speech signal, and how can they be implemented in an electromechanical improvisation and composition practice?
- R2. How can NAS models trained on speech datasets be incorporated into this practice, and what methods can be used to organise generated material?
- R3. What are the affordances and limitations of text models as creative tools, applied to linguistic speech structures?

These questions are addressed through the following objectives:

- 1. Conduct a field review of existing computational methods in speech-derived music to identify common methods and gaps in composition and sound art practices, including but not limited to electromechanical instruments.
- 2. Conduct a field review of machine learning applications in electromechanical instruments to identify methods and gaps within broader electromechanical practice, including but not limited to speech-derived music.

46. Corey Ford et al., ‘Reflection Across AI-based Music Composition’, in *Proceedings of the 16th Conference on Creativity & Cognition* (2024), 398–412, <https://doi.org/10.1145/3635636.3656185>.

47. Pedro Ferreira, Ricardo Limongi and Luiz Paulo Fávero, ‘Generating music with data: application of deep learning models for symbolic music composition’, *Applied Sciences* 13, no. 7 (2023): 4543, <https://doi.org/10.3390/app13074543>.

48. Şöl en Kıratlı, Hannah E Wolfe and Alex Bundy, ‘Cacophonic choir: an interactive art installation embodying the voices of sexual assault survivors’, in *ACM SIGGRAPH 2020 Art Gallery* (Association for Computing Machinery, 2020), 446–450, <https://doi.org/10.1145/3386567>.

49. Jennifer Walshe, *The Text Score Dataset 1.0*, 2021, accessed 25 February 2026, <http://milker.org/s/The-Text-Score-Dataset-10-Jennifer-Walshe.pdf>.

3. Develop compositional methods employing computational speech feature extraction, NAS-generated audio, LSTM and SML generated text in response to findings from each field review.
4. Apply transcription methods and build two DIY motor-driven devices for the mechanical realisation of speech audio.
5. Train RAVE audio models and build two DIY motor-driven devices for the realisation of compositions incorporating NAS-generated audio with electromechanical elements.
6. Train text models and build three DIY motor-driven devices for installation, improvisation and composition around generated linguistic patterns.
7. Develop an improvisatory and compositional practice through performance and composition with each device.
8. Reflect on the practical implementation of the methods and devices as captured in the portfolio, through a commentary which engages critically with the creative process.
9. Contextualise the creative output of this research in the fields of speech-derived and electromechanical music, and the broader field of practice research in contemporary music and sound art.

1.7 Methodology

To address each question, this research is carried out in two parts: first, field reviews to identify current methods, trends, and gaps. Second, the creation of portfolio of creative outputs, each informed by and developing methods identified in the field reviews. Creative works are evaluated through a mixed method approach, using qualitative measures and critical reflection. Qualitative measures allow me to document the behaviour of the devices and models, while critical reflection is necessary to explain the compositional choices I made in response to those behaviours. This reflection aims to clarify insights emerging from the practical activity of developing new work—what Barbara Bolt describes as a ‘reflexive knowing that imbricates and follows on from handling’.⁵⁰

50. Barbara Bolt, ‘The Magic is in Handling’, in *Practice as research: Approaches to creative arts enquiry*, ed. Estelle Barrett and Barbara Bolt (Ib Tauris London, 2007), 34, <https://doi.org/10.5040/9780755604104.ch-002>.

1.7.1 Field Review

The first field review, Computational Methods in Speech-Derived Music (Chapter 2), looks at contemporary composition practices using speech as a source of musical material. I survey methods, speech sources, instrumentations, and shared themes, showing a growing trend towards live processing and speech resynthesis.

A second field review, Machine Learning in Electromechanical Sound Art (Chapter 3), looks specifically at the integration of neural networks within electromechanical works, focusing on questions of material constraints and approaches to interaction. While there are relatively few examples of speech-derived mechanical music, the methods and conceptual approaches identified here influence my creative practice throughout the submitted portfolio.

1.7.2 Speech Sources

Informed by the range of speech sources found in the field reviews, each creative project uses a different source. These include recorded speech, generated speech audio, and generated text. Recorded speech covers individual speakers, datasets of emotional speech, and audiobooks. Generated speech is produced using RAVE models; training data included whispered speech, overlapping speech, emotional speech, and words scraped from online dictionaries. Generated text is produced using LSTMs and small language models, using datasets of dialogue acts, dependency relations, and word embeddings.

1.7.3 Feature Selection and Mapping

As indicated above, a range of speech sources are explored throughout this project. The choice of a particular speech feature, some of which are more abstracted than others, is a crucial part of the compositional process. Several subjective criteria inform my decisions: how straightforward it is to extract a particular feature, which qualities are useful musically—such as range or variation—or more practical concerns, such as whether the data is continuous or discrete, how this fits into a mapping process, and the constraints of a physical device.

The importance of choices regarding data representation can be seen in Ablinger's *Quadraturen III*, where the representation of speech through player piano is a priority: the work depends on the 'border of recognition' within reinterpretation.⁵¹ If however our concerns are

51. Ablinger, *Quadraturen III*.

tilted towards the musical—that is, focused on the aesthetic qualities of the result—then choices will be made which obscure the connection between a speech source and consequent material. In this sense, there is a tension between the use of speech as a compositional starting point, the aesthetic qualities of the output, and the transparency of the process which connects them.

This concern occurs across algorithmic composition, as Agostino Di Scipio describes his mapping approach to Paul Doornbusch:

“It’s the emergent properties ... which are the very object of the mapping, which should inform the audible phenomenon, be it extended musical structure or single sound event.”⁵²

Doornbusch himself describes music as ‘the main goal ... I will intervene in the [mapping] process to achieve the aesthetic result I want’. A flexible approach to the musical qualities of a source can be found in Friedmann and Schmucki’s approach to speech, describing ‘sprache as mutable content, as a changeable structure’.⁵³

1.7.4 Devices

Having considered the speech sources and mapping process, this section turns to the DIY electromechanical devices to which these sources and features can be mapped. Speech data can be reproduced digitally with total transparency, but a physical device introduces material constraints; a weight and friction that can be entirely unrelated to the human voice. This diverges from the historical goals of emulating human speech⁵⁴ to focus on an intentional mismatch between the complexity of speech and the simplicity of a motor-driven device. To me, the DIY electromechanical approach allows an exploration of materiality as a compositional constraint—a topic of ongoing relevance in contemporary sound art.⁵⁵

I see the value of this approach in the unpredictable interaction between speech data and physical materials. There is no requirement for the objects to be drawn from speech, or designed to imitate any aspect of human vocal production—in my view, it is only through the difference

52. Doornbusch, ‘Composers’ views on mapping in algorithmic composition’.

53. Annette Schmucki and Alistair Zaldúa, ‘Music and Language Tilt into Each Other’: Annette Schmucki Interviewed by Alistair Zaldúa’, *Contemporary Music Review* 39, no. 3 (2020): 373–385, <https://doi.org/10.1080/07494467.2020.1821528>.

54. Ondrejovic, ‘Wolfgang von Kempelen and His Speaking Machine’; Helmholtz, ‘XII. On vowel sounds’.

55. Valle, ‘Making acoustic computer music: The Rumentarium project’; Zareei, ‘Audiovisual materialism’; Flø, ‘Materiality in sound art’; Pigott, ‘Across Fields: Sound, art and technology from an electromechanical perspective’.

between the source and material realisation that new and surprising musical material emerges. The use of found and recycled objects is an important aspect; as with the use of speech, this involves placing ordinary objects in a musical context, with their imperfections and limitations drawn out through the complexity of speech mappings. These devices cannot be separated from the sources and methods used. They are the direct result of improvisation with datasets, mapping methods, mechanisms and found objects. Each was developed iteratively through practice, and employ different combinations of electric motors, boards, and microcontrollers.

1.7.5 Improvisation

Musical improvisation throughout this project allowed the physical realisation of each method through the material constraints of each device, which I engage with in a performative artistic practice. This approach can be found in Andrea Valle's description of a focus on limited, lo-fi instruments as a 'reductionist' approach in which instruments are designed, developed and performed through improvisation.⁵⁶ In the first case, the construction of each device is a form of improvisation with found materials and DIY electronics, and a recurring practice in DIY sound-making.⁵⁷ This is related to Bowers and Archers idea of infra-instruments,⁵⁸ in which deliberately limited musical instruments are often constructed from materials to hand.

Improvisation extends to the control software for each device and into live performance, with written compositions also using musical material initially developed through improvisation. Improvisation here affords a tactile relationship with both the computational methods employed, and the physical materials of each DIY device. I find that this approach allows a thorough exploration of each system, as well as the discovery of unexpected behaviours that emerge through performance, rather than design.⁵⁹

1.8 Contributions

The main contributions of this research are:

56. Valle, 'Making acoustic computer music: The Rumentarium project'.

57. John Richards and Tim Shaw, 'Improvisation through Performance-installation', *Organised Sound* 27, no. 2 (2022): 144–155, <https://doi.org/10.1017/S1355771821000583>.

58. John Bowers and Phil Archer, 'Not hyper, not meta, not cyber but infra-instruments', in *Proceedings of the International Conference on New Interfaces for Musical Expression* (2005), 5–10, <https://doi.org/10.5281/zenodo.1176713>.

59. Elizabeth Hallam and Tim Ingold, *Creativity and cultural improvisation* (Routledge, 2021), 3–15.

- Field reviews providing an overview of current practice, and context for this research.
- A set of compositional strategies for mapping speech and language to DIY electromechanical devices, outlined in Chapter 4, 5 and 6 of this commentary.
- A portfolio of original devices, compositions, scores, and recordings applying the strategies described.
- Publicly-available code for each compositional strategy, in the *Code* folder, and online.⁶⁰
- Trained RAVE and text models developed through this research, in the *Models* folder, and online.⁶¹

1.9 Overview of Contents

Portfolio

The *Portfolio* contains recordings of the creative output from this research, with scores and examples where relevant. Folders are organised by chapter and creative project.

- **Chapter 4 - Motorcase I**

Motorcase I is a DIY motor-driven device for improvisation around a small set of speech audio transcriptions, with speech signal-extracted features including onsets, the spectral centroid, and rhythmic segmentation derived from text transcription. Duo improvisation, video composition.

- **Examples/...** contains recordings of each object used, and examples of specific features of the process.
- **Motorcase 1 improvisation.mov** presents an improvisation with the device in collaboration with Gaia Blandina (cello).
- **Years a have to know or something.mov** composed for small ensemble, device, and video, notated in **Years a have to know or something - score.pdf**. Sample-based recording.

- **Chapter 4 - Motorcase II**

Motorcase II takes an emotional speech corpus as the basis for a larger motor-driven

60. https://github.com/fifoh/-Speech_Derived_Music_Code

61. <https://huggingface.co/fifoh/speech-derived-music-models/tree/main>

improvisation system, navigating the dataset through emotion labels. Duo EP, solo improvisation, composition for ensemble and tape.

- **Examples/...** contains an example of the transcription method in comparison with the original speech audio.
- **Duo EP/...** contains four tracks of improvised music in collaboration with Gaia Blandina (cello). Recorded by Jack McNeill.
- **Solo improvisation.mov** presents a video of improvisation with the device.
- **Machine Translation.wav** composed for small ensemble, device, and tape, notated in **Machine Translation - score.pdf**. Played by Danny Saleeb (violin), Bethany Nichol (bass clarinet), James Wood (drums). Recorded by Alex Mackay.
- **Preamble/...** contains video of a precursor to *Motorcase II*.

• Chapter 5 - Smalltalk

Smalltalk is a fixed composition for three devices, incorporating NAS-generated audio, and selecting material using a method of pitch-contour extraction.

- **Examples/...** contains two folders with audio examples relevant to the compositional approach.
- **SmalltalkVideo.mov** presents a video of the composed work for three devices.

• Chapter 5 - Radio User Guide

Radio User Guide is written for three devices built partially from discarded radio and CD-player parts, using RAVE models of speech and music trained to different levels, and organising generated material using a text model trained on radio manuals. Fixed composition.

- **Radio_User_Guide_audio.wav** composed for three devices, notated in **Radio User Guide - score.pdf**.
- **Radio User Guide.mov** presents a video extract of the composed work. Played by Danny Saleeb, Ben Maloney (MIDI keyboards).
- **Examples/...** contains audio recordings showing the application of particular compositional strategies.

- **Chapter 6 - Porcelain Music Box**

Porcelain Music Box is a small scale installation work which takes as a starting point a dataset of dialogue acts. An LSTM model trained on the dataset is used to trigger algorithmically generated material, in patterns of continuation and disruption.

- `Porcelain Music Box.mov` presents a short video of the installation work.

- **Chapter 6 - Seven Studies for Electric Motors**

Seven Studies for Electric Motors is a modular mechanical device, with its output determined by dependency structures produced by an SLM. Improvisation.

- This folder contains seven videos; each study of the work. Filmed by Bob Tsang.

- **Chapter 6 - Word games**

Word games is a modular mechanical device using fixed text to generate musical sequences via word embeddings, overlaying subtitles of the original text on live video of the device. Composition.

- `wordgames.mov` presents a video of the work.

Code

The *Code* folder contains scripts for each method and device developed throughout this practice research.

- **Chapter 4 - Motorcase I code:**

- `TranscriptionMethod.py` for converting speech audio to motors.
- `ImprovPatch.maxpat` for live improvisation.
- `Video composition patch.maxpat` for performing the video piece.

- **Chapter 4 - Motorcase II code:**

- `Transcription.py` for transcribing a dataset of speech audio to MIDI.
- `Improvisation.py` for live improvisation, `CircuitPy` for the motors, and other relevant files.

- **Chapter 5 - Smalltalk code:**

- `MatchingGestures.py` for identifying pitch contours in audio, with `MatchingGesturesLongPlot.py` for longer duration files.
- `SmalltalkSEND.py` and `SmalltalkRECEIVE.py` for performing the precomposed work.
- **Chapter 5 - Radio User Guide code:**
 - `radio model DEMO.py` for running the text model.
 - `performance.py` for performing the composition.
- **Code/Chapter 6 - Porcelain Music Box code:**
 - `Porcelain Music Box code.py` for running the installation piece.
- **Chapter 6 - Seven Studies for Electric Motors code:**
 - `pd_patch.pd` for coordinating text, user input, LED's and motors.
 - `generate_sentences.py` for text generation using an existing Small Language Model.
 - `LED_receive.txt` for lights, and `OSC_recieve.py` for motors.
- **Chapter 6 - Wordgames code:**
 - `words5.py` for live performance.
 - `camera_7.py` for overlaying text subtitles.
 - Additional files for the reduced word embeddings described in the commentary.

Models

The *Models* folder contains LSTM text models, and RAVE audio models trained for each creative project.

- **Chapter 4 - Motorcase II models**
 - This folder contains four RAVE models, each trained on a different region of an emotional speech dataset. `info.txt` contains the training configuration.
- **Chapter 5 - Radio User Guide models**
 - `RAVE models/...` contains folders for each model, trained to varying degrees, with `allWordsAudio.wav` as the training dataset for the speech RAVE model. Files containing the training configuration are included.

- **Text models/...** contains .onnx files for each training stage of the text model, and other necessary files. **Radio manuals augmented.txt** is the training dataset.
- **Chapter 5 - Smalltalk models**
 - This folder contains RAVE models trained on ordinary, whispered, and overlapping speech, as well as training scripts.
- **Chapter 6 - Porcelain Music Box models**
 - This folder contains **LSTM_words_BASIC_6voices.tflite** for running the dialogue acts model, and other necessary files.

Written Commentary

- **Chapter 2** presents a field review of computational methods in speech-derived music.
- **Chapter 3** presents a field review of machine learning applications in electromechanical sound art.
- **Chapter 4** explores algorithmic transcription methods, through two creative works:
 - *Motorcase I*
 - *Motorcase II*
- **Chapter 5** discusses the incorporation of audio models, through two creative works:
 - *Smalltalk*
 - *Radio User Guide*
- **Chapter 6** looks at the use of text models for structural organisation, through three creative works:
 - *Porcelain Music Box*
 - *Seven Studies for Electric Motors*
 - *Word games*
- **Chapter 7** presents conclusions, revisits the research questions, and suggests directions for future research.

1.10 Conclusion

In this chapter I have outlined the basis of this research, introducing the motivation to derive musical material from speech and realise this material through DIY motor-driven devices. Research questions have been presented, addressing how algorithmic transcription, RAVE audio models, and text models might each contribute to a practice of speech-derived electromechanical sound art.

A methodology was defined in two stages: field reviews, followed by a creative practice combining feature analysis, mapping, and iterative design through improvisation and experimentation. The structure of this creative practice was outlined, and the overall contributions of this research have been listed.

In the following chapter, I examine existing computational methods in speech-derived music, to provide an overview of the current field and context for the practice that follows.

Chapter 2

Computational Methods in Speech-Derived Music

Recent years have seen the emergence of diverse computational methods used to derive musical material from speech. While speech analysis, synthesis, and signal processing advancements have informed contemporary music works, a review of computational methods in speech-derived music is absent. To begin to address R1—on which features can be computationally extracted from the speech signal, and their implementation in practice—this practice review addresses the gap, categorising methods from 1980 to 2023 found in literature and contemporary music festival programs. The chapter identifies trends relating to live computation, automated compositional processes, use of acoustic-mechanical instruments, and themes including human-machine interaction, obscuring of semantic meaning, and highlighting geographical speech qualities. These approaches have broadened compositional palettes and fostered novel composer-performer-electronic interactions, which inform the creative output of Chapters 4, 5 and 6.

2.1 Introduction

The use of speech as musical source material is not a new concept. While speech in contemporary musical composition is often associated with the works of Steve Reich,¹ the musical notation of speech can be found in many other examples, including but not limited to Steele's 18th-

1. Eric Prieto, 'Speech Melody and the Evolution of the Minimalist Aesthetic in Steve Reich's *The Cave*', *Circuit* 12, no. 2 (2002): 21–44, <https://doi.org/10.7202/902250ar>.

century transcriptions of recited Shakespeare,² the early twentieth-century transcriptions of Leoš Janáček,³ and the development of vocal sampling through the musique concrète works of Pierre Schaeffer.⁴

Speech-derived works cover a range of fields including speech sampling, mimicking speech with instruments and electronics, and synthesised speech. Computational methods have been developed by several composers in recent years with advances in analysis, synthesis, and signal processing influencing practice from score preparation to live improvisation.⁵

Computational methods include, for example, Philip Venables's use of piano key-stroke detection to trigger speech audio in *Answer Machine Tape, 1987* (2022), Jonathan Harvey's use of automated transcription to generate a musical score in *Speakings* (2007), or Trevor Wishart's interpolation between speech audio fragments and other sounds in *Vox 5* (1986).

Despite consistent artistic interest in this area, a review of computational methods in speech-derived music is lacking. Related work includes Lane's review of compositional approaches to recorded speech⁶ and a systematic review of voice at the New Interfaces for Musical Expression conference.⁷ This study draws on Lane's work but differs in compiling and categorising works according to computational methods, rather than source material or transmission of meaning. The methodology section of this chapter explains how works were identified and categorised, followed by a presentation of selected works. The discussion of this chapter includes a summary of methods, sound sources, artistic intent, instrumentation, challenges, and trends, and shows a growing trend towards live processing and resynthesis.

The methods presented in this chapter are connected through speech as source material and a computational aspect in the compositional process. These methods have expanded the palette of sounds and structures available to composers and facilitated new forms of interaction among

2. Noelle Chao, 'Listening to the Voice on the Page: Joshua Steele and Technologies of Recording', *The Eighteenth Century* 54, no. 2 (2013): 245–261, <https://doi.org/10.1353/ecy.2013.0012>.

3. Paul Wingfield, 'Janáček's speech-melody theory in concept and practice', *Cambridge Opera Journal* 4, no. 3 (1992): 281–301, <https://doi.org/10.1017/s0954586700003803>.

4. Thom Holmes, *Electronic and experimental music: technology, music, and culture* (Routledge, 2012), 55.

5. Lorenzo Cardilli, Stefano Lombardi Vallauri et al., *L'arte orale. Poesia, musica, performance* (Accademia University Press, 2020), <https://doi.org/10.4000/books.aaccademia.7908>; Leo Chadburn, 'Captions, characters, self-portraits: compositional approaches to the disembodied speaking voice and the voice-text-music relationship' (PhD diss., City, University of London, 2023); Lauren Redhead, 'Language as a Material Process', *Contemporary Music Review* 39, no. 3 (2020): 359–372, <https://doi.org/10.1080/07494467.2020.1821527>.

6. Cathy Lane, 'Voices from the Past: Compositional approaches to using recorded speech', *Organised Sound* 11, no. 1 (2006): 3–11, <https://doi.org/10.1017/s1355771806000021>.

7. Rébecca Kleinberger et al., 'Voice at NIME: a Taxonomy of New Interfaces for Vocal Musical Expression', in *Proceedings of the International Conference on New Interfaces for Musical Expression* (2022), <https://doi.org/10.21428/92fbeb44.4308fb94>.

composers, performers, and electronics.

2.2 Methodology

I searched research databases, contemporary music festival programs, and archives to cover academic and practice-based methods. Dates were chosen to limit the scope of the review. The information sources were:

- RILM Abstracts of Musical Literature: ‘composition AND speech AND derived NOT (classification or categorisation or discrimination)’; ‘composition AND speech AND computer NOT (classification or categorisation or discrimination)’; ‘composition AND speech AND method NOT (classification or categorisation or discrimination)’
- New Interfaces for Musical Expression proceedings: ‘speech, speak, spoke, language, syllable, phoneme, phonetic, conversation, diction, dialect, vowel, consonant, transcribe, transcript’
- Huddersfield Contemporary Music Festival programs (search terms above)
- Donaueschinger Musiktage archive: composers with search term ‘speech’

Sources that returned duplicates included IEEE Xplore, Computer Music Journal, ICMC Proceedings, Wein Modern, Darmstadt Summer Festival, and Warsaw Autumn festival archives.

Initial results were cut based on inclusion and exclusion criteria:

- Literature and compositions published/written from 1980 to 2023
- Using a computational method: intuitive pitch transcription of speech alone was not relevant
- Singing voice synthesis results were removed
- Classification, categorisation, or discrimination of speech results were removed

In most cases, a single work by a composer was included. Multiple works were included if different methods were used.

2.3 Results

The literature search returned two field reviews, five overview articles (exploring multiple composers or a specific method), twelve journal articles, six conference papers, and three theses. Contemporary classical music programs and relevant works described in the literature returned forty-five musical works.

The results were categorised according to computational methods relating to the treatment of speech. Categories and subcategories were determined by hand. In Table 2.1, a single prominent method has been identified in each work.

Table 2.1: Computational methods, categories and works.

Category	Subcategory	Works
Sample playback	Fixed tape tracks	[8], [9], [10], [11]
	Live triggering	[12], [13], [14], [15]
Electronic manipulation	Filtering	[16], [17], [18], [19]
	Delays and reverb	[20]
	Loops and repetition	[21], [22], [23], [24]
	Granulation	[25]
	Interpolation	[26], [27], [28]
	Altering speed	[29], [30], [31]
	Pitch shifting	[32]
Mapping	Instrumental resynthesis	[33], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47], [48]
	Text/Visual mapping	[49], [50], [51], [52], [53], [54], [55]
Speech synthesis	Linear predictive coding	[56], [57]
	Markov models	[58]
	Neural synthesis	[59]

WORKS IN TABLE 2.1:

- [8] François Sarhan, *L’Nfer*, 2012, accessed 25 February 2026, <https://francoissarhan.blogspot.com/search?q=L%27Nfer>
- [9] Henrik Hellstenius, *Instrument of Speech*, 2016, accessed 25 February 2026, https://hellstenius.no/wp-content/uploads/2018/09/PLING_An-expanded-palette-2.pdf
- [10] Alessandro Bosetti, *Alpine Flipbooks*, 2013, accessed 25 February 2026, <https://melgun.net/travaux/alpine-flipbooks/>
- [11] Alessandro Bosetti, *The Notebooks*, 2014, accessed 25 February 2026, <https://melgun.net/travaux/the-notebooks/>
- [12] Matthew Grouse, *Left Right, Left Right*, 2020, accessed 25 February 2026, <https://www.matthewgrousemusic.com/left-right-left-right>
- [13] Phillip Venables, *Answer Machine Tape, 1987*, 2022, accessed 25 February 2026, <https://philipvenables.com/2022/09/25/answer-machine-tape-1987>
- [14] Ondrej Adámek, *Ça tourne ça bloque*, 2007, accessed 25 February 2026, <https://ondrejadamek.com/listen/ca-tourne-ca-bloque-10-musiciansamplervideo/>
- [15] Alessandro Bosetti, *Mask Mirror*, 2008, accessed 25 February 2026, <https://melgun.net/travaux/maskmirror/>
- [16] Paul Lansky, *Word Color*, 1992, accessed 25 February 2026, https://paul.mycpanel.princeton.edu/liner_notes/morethanidlechatter.html#
- [17] Hector Parra, *Fragments on Fragility*, 2009, accessed 25 February 2026, <https://hectorparra.net/docs/program-notes/>
- [18] Cecilie Ore, *Etapper*, 1988, accessed 25 February 2026, <https://cecilieore.no/node/92>
- [19] Freya Bailes and Roger T Dean, ‘When is noise speech? A survey in sonic ambiguity’, *Computer Music Journal* 33, no. 1 (2009): 57–67, <https://doi.org/10.1162/comj.2009.33.1.57>
- [20] Lauren Redhead, *Séo níedhæmestre; se tidfara*, 2018, accessed 25 February 2026, <https://panyrosasdiscos.bandcamp.com/track/s-o-n-edh-mestre-se-tidfara>
- [21] Pamela Z, *Geekspeak*, 1995, accessed 25 February 2026, <https://pamelaz.com/work/geekspeak/>
- [22] Lottie Sadd, *To Copy the Speech without Saying It Directly*, 2021, accessed 25 February 2026, <https://thebookroom.net/here-here-on-voice-and-electronics-territorialise-deterritorialise-reterritorialise/>
- [23] Bernhard Gál, *bestimmung new york*, 1999, accessed 25 February 2026, <https://bernhardgal.com/bestimmungnewyork.html#info>
- [24] Blablabor, *hirsch hirn hornisse*, 2009, accessed 25 February 2026, <https://www.blablabor.ch/hirsch-hirn-hornisse>
- [25] Barry Truax, *Pendlerdrøm*, 1997, accessed 25 February 2026, <https://electrocd.com/en/oeuvre/13157/barry-truax/pendlerdrøm>
- [26] Trevor Wishart, *Vox 5*, 1986, accessed 25 February 2026, <https://britishmusiccollection.org.uk/score/vox-5>
- [27] Wende Bartley, *Ellipsis*, 1988, accessed 25 February 2026, <https://electrocd.com/en/oeuvre/14175/wende-bartley/ellipsis>
- [28] Bruno Ruviano, *The Spell of Speech*, 2004, accessed 25 February 2026, <https://scholarcommons.scu.edu/music/20>
- [29] Clarence Barlow, *fLvXvS*, 1990, accessed 25 February 2026, <https://clarlow.org/compositions-by-year>
- [30] Joanna Bailie, *Artificial Environments no.8*, 2013, accessed 25 February 2026, <https://joannabailie.com/music/solo-and-small-chamber/ae8/>
- [31] Trevor Wishart, *Globalalia*, 2004, accessed 25 February 2026, <https://www.soundohm.com/product/globalalia-imago>
- [32] John Young, *An Angel at Mons*, 2014, accessed 25 February 2026, <https://empreintesdigitales.bandcamp.com/track/an-angel-at-mons>
- [33] Clarence Barlow, *Orchideæ Ordinariæ*, 1989, accessed 25 February 2026, <https://www.soundohm.com/product/musica-algorithmica-2-cd>
- [34] Jonathan Harvey, *Speakings*, 2007, accessed 25 February 2026, <https://www.fabermusic.com/shop/speakings-p410799>
- [35] Thomas Hummel, *Nicanor*, 1996, accessed 25 February 2026, <https://www.thomashummel.net/en/music/nicanor/>
- [36] Andrew Chen, ‘Voice Without Speaker: human speech synthesis in acoustic instrumental contexts’, *Tempo* 77, no. 305

(2023): 60–71, <https://doi.org/10.1017/s0040298223000098>

[37] Beat Furrer, *Wüstenbuch*, 2009, accessed 25 February 2026, <https://www.kairos-music.com/cds/0013312kai>

[38] Luke Lewis, *The Echoes Return Slow*, 2021, accessed 25 February 2026, <https://luke-lewis.com/music/theechoesreturnslow.html>

[39] Claudy Malherbe, *Locus*, 1997, accessed 25 February 2026, <https://brahms.ircam.fr/en/works/work/10464/>

[40] Peter Ablinger, *Voices and Piano*, 1998–2003, accessed 25 February 2026, https://ablinger.mur.at/docs/v+p_ginsburgh.pdf

[41] Clarence Barlow, *Farting Quietly in Church*, 1992, accessed 25 February 2026, <https://clarlow.org/compositions-by-year>

[42] Mauro Lanza, *The Kempelen Machine*, 2015, accessed 25 February 2026, <https://brahms4.ircam.fr/fr/works/work/41088/>

[43] Martin Riches, *Talking Machine*, 1989, accessed 25 February 2026, <https://martinriches.de/>

[44] Peter Ablinger, *Quadraturen II*, 2002, accessed 25 February 2026, <https://ablinger.mur.at/docu11.html#qu2>

[45] Fou Yoshimura and Kazuhiro Jo, ‘A "voice" instrument based on vocal tract models by using soft material for a 3D printer and an electrolarynx.’, in *Proceedings of the International Conference on New Interfaces for Musical Expression* (2019), 411–412, <https://doi.org/10.5281/zenodo.3673015>

[46] Charles Holbrow, Elena Jessop and Rébecca Kleinberger, ‘Vocal vibrations: A Multisensory Experience of the Voice’, in *Proceedings of the International Conference on New Interfaces for Musical Expression* (2014), 431–434, <https://doi.org/10.5281/zenodo.1178800>

[47] Florian Hecker, *Inspection II*, 2019, accessed 25 February 2026, <https://lampo.org/archive/florian-hecker-2019/>

[48] Daniel Buner Formo, ‘The Orchestra of Speech—a speech-based instrument system’, in *Proceedings of the International Conference on New Interfaces for Musical Expression* (2018), <https://doi.org/10.5281/zenodo.1302607>

[49] J. Vassallo, *Versificator*, 2021, accessed 25 February 2026, <https://nivel.teak.fi/carpa7/versificator-algorithmic-poetry-and-music-composition/>

[50] Timothy D Polashek, ‘Beyond Babble: A Text-Generation Method and Computer Program for Composing Text, Music and Poetry’, *Leonardo Music Journal* 15 (2005): 17–22, <https://doi.org/10.1162/lmj.2005.15.1.17>

[51] Clarence Barlow, *Progéthal Percussion for the Advanced Beginners*, 2003, accessed 25 February 2026, <https://clarlow.org/compositions-by-year/>

[52] Clemens Gadenstätter, *Semantical Investigations II*, 2007, accessed 25 February 2026, <https://www.kairos-music.com/cds/0015006kai>

[53] Harald Muenz, *deChiffreAGE*, 2004, accessed 25 February 2026, <https://www.musicwedliketohear.com/pdf/29032006.pdf>

[54] Jaap Blonk, *Traces of Speech*, 2014, accessed 25 February 2026, <https://jaapblonk.bandcamp.com/album/traces-of-speech>

[55] Helmut Oehring, *[im'ferno] (from: MAPPA) Contrapasso I – V (to: Wladimir Putin / Sergej Lawrow)*, 2022, accessed 25 February 2026, <https://helmutoehring.de/kopie-von-2021-22>

[56] Paul DeMarinis, *Alien Voices*, 1988, accessed 25 February 2026, <https://pauldemarinis.org/AlienVoices.html>

[57] Paul DeMarinis, *Music as a Second Language*, 1991, accessed 25 February 2026, <https://soundohm.com/product/music-as-a-second-language-2>

[58] Maria Astrinaki et al., ‘MAGE 2.0: New Features and its Application in the Development of a Talking Guitar’, in *Proceedings of the International Conference on New Interfaces for Musical Expression* (2013), 547–550, <https://doi.org/10.5281/zenodo.1178467>

[59] Memo Akten and Jennifer Walshe, *ULTRACHUNK*, 2018, accessed 25 February 2026, <https://memo.tv/works/ultrachunk/>

2.4 Computational Methods: Selected Works

The following sections examine a single work from each subcategory to illustrate the specific technical processes and creative applications that define each method.

2.4.1 Sample Playback

Seven works were found where speech samples were a prominent compositional element and had not been significantly processed. Three used fixed tape tracks, and four used live triggering. While sampling occurred as a stage in most methods, later examples look at other primary methods.

Fixed Tape Tracks

The use of pre-recorded speech in contemporary music composition is perhaps most associated with the music of Steve Reich, but this method of speech-melody composition developed by Scott Johnson⁶⁰ is one of many ways composers have used fixed tape tracks of speech. These works included intuitive instrumental transcription, live narration, and video.

- Alessandro Bosetti's video piece *Alpine Flipbooks*⁶¹ uses samples of specific words recorded from the French-Italian border and Slovenia. A Max-MSP (Cycling '74) patch plays different versions of each word, each paired with a distinctive dance, making visible the gradual and abrupt changes in dialect and language across a geographical region.

Live Triggering

Four works were identified in which live triggering of samples was a central element. Three used predetermined samples; in these, the effect was like the use of a fixed tape track. Works included percussion, piano, and custom software instruments triggering pre-recorded speech samples. The dates suggest a recent trend towards live triggering rather than fixed tape tracks, affording players and ensembles flexibility in performance.

- Philip Venables's *Answer Machine Tape, 1987*⁶² takes the answering machine messages

60. Jonathan Maltman and Lindsay Vickery, 'The Foundations of Speech-Melody Composition And The Dawn Of The Digital Music Era', in *Proceedings of the Australasian Computer Music Conference* (2018), 30–37.

61. Bosetti, *Alpine Flipbooks*.

62. Venables, *Answer Machine Tape, 1987*.

of David Wojnarowicz. A novel piano key-stroke detection method triggers the speech sounds, treating the piano as a typewriter to “transcribe, comment on and illuminate the messages” in an exploration of queer history and the AIDS crisis.

2.4.2 Electronic Manipulation

Seventeen examples used electronic speech manipulation as a central part of the work. These methods are widely used in works unrelated to speech but are given new purpose in this context, obscuring or highlighting semantic meaning and drawing out inherent musical qualities of speech.

Filtering

This category covered a range of techniques, including filtering the speech signal, noise, and instrumental sound.

- Hector Parra’s *Fragments on Fragility*⁶³ applies fricative consonantal filters to a live string quartet, producing string sounds with a perceptual similarity to speech in an exploration of the relationship between ‘glottal impulses of the human voice and the vibratory states of the string instruments.’

Delays and Reverb

The use of analog and digital delay and reverb were used in several works, though rarely as the main element.

- Lauren Redhead’s *Séio níedhæmestre; se tidfara*⁶⁴ for chamber organ and electronics uses Gavin Bone’s translations of the poems *The Seafarer*, *The Wanderer*, and others (1950). Reverb, delays, and ring modulation were used to emphasise audio artifacts and errors and to present the ‘possibility of sense-making whilst simultaneously denying it’.⁶⁵

Loops and Repetition

Common techniques included the layering of looped samples—rhythmic looping in which semantic content is rendered meaningless—and structures that return to the same repeated sound.

63. Parra, *Fragments on Fragility*.

64. Redhead, *Séio níedhæmestre; se tidfara*.

65. Redhead, ‘Language as a Material Process’, 9.

- Lottie Sadd's *to copy the speech without saying it directly*⁶⁶ for voice and live electronics uses the performers' live speech in a semi-improvised duet exploring live repetition and imitation, in which speech is 'rendered meaningless in its mimicry.'

Granulation

- Barry Truax's *Pendlerdrøm*⁶⁷ for tape applied granulation to train station announcements, combined with resonators and time stretching of the sound samples. The aim was to bring out the character of source sounds, make them more 'musically defined,' and evoke the experience of daydreaming.⁶⁸

Interpolation or Combination

There were three works in which a prominent method involved combining speech sounds or combining speech with other sounds:

- Trevor Wishart's *Vox 5*⁶⁹ for tape takes a recording of a human voice speaking and using extended vocal techniques (i.e., ingressive speech, subharmonics, ululation, special consonant production). Interpolatable sounds were identified by preprocessing, analysing, and resynthesising the sources using a phase vocoder, and using programs for spectral expansion and contraction. Perceptual boundaries between sources informed musical decision-making, with an aesthetic goal of maintaining clarity of musical discourse through the transformation of speech into other recognisable sound sources.⁷⁰

Altering Speed or Direction

- Joanna Bailie's *Artificial Environments no.8*⁷¹ for piano and tape uses a freezing technique in the movement 'Babel'; sounds were sourced from field recording excerpts of a queue outside Notre Dame. The freezing technique forms harmonic progressions from which the piano material is drawn.

66. Sadd, *To Copy the Speech without Saying It Directly*.

67. Truax, *Pendlerdrøm*.

68. Barry Truax, 'Speech, music, soundscape and listening: interdisciplinary explorations', *Interdisciplinary Science Reviews* 47, no. 2 (2022): 279–293, <https://doi.org/10.1080/03080188.2022.2035103>.

69. Wishart, *Vox 5*.

70. Trevor Wishart, 'The Composition of "Vox-5"', *Computer Music Journal* 12, no. 4 (1988): 21–27, <https://doi.org/10.2307/3680150>.

71. Bailie, *Artificial Environments no.8*.

Pitch Shifting

- John Young's *An Angel at Mons*⁷² for tape takes a BBC archive recording of John Ewings speaking about an experience in the First World War. Frequency shifting and envelope shaping of noise are used to approximate the sibilance of Ewings' speech and create timbrally related material. The aim was to accentuate nonverbal speech sonorities, as a sonic representation of the speaker beyond recorded speech.⁷³

2.4.3 Mapping

Instrumental Resynthesis

This section looks at works using instrumental resynthesis or mapping as a primary method, including electronic works and sound-making machines. Many used spectral analysis; common programs included OpenMusic (partial tracking), SPEAR (sine wave resynthesis), and bespoke programs developed by composers. Recent works showed tendencies towards automated orchestration methods and wider variation in the parameters of audio analysis.

- Jonathan Harvey's *Speakings*⁷⁴ for orchestra and live electronics uses recordings of radio interviews, baby babbles, and poetry readings. An automated orchestration method was developed using Orchidée⁷⁵ to match speech dynamics, brightness, and roughness to features of instrumental sound using an evolutionary algorithm. Live filtering is used to approximate speech formants, creating the impression of an orchestra 'learning to speak'.⁷⁶
- Claudy Malherbe's *Locus*⁷⁷ for vocal quartet and electronics uses fragments taken from technical works about vocal acoustics, phonetics, and phonology. The work explores a process of reduction through automated transcription, using AudioSculpt⁷⁸ and OpenMusic⁷⁹

72. Young, *An Angel at Mons*.

73. John Young, 'Figures of Speech: Oral History as an Agent of Form in Electroacoustic Music', *Leonardo Music Journal* 28 (2018): 88–94, https://doi.org/10.1162/lmj_a_01047.

74. Harvey, *Speakings*.

75. Grégoire Carpentier, Gérard Assayag and Emmanuel Saint-James, 'Solving the musical orchestration problem using multiobjective constrained optimization with a genetic local search approach', *Journal of Heuristics* 16, no. 5 (2010): 681–714, <https://doi.org/10.1007/s10732-009-9113-7>.

76. Gilbert Nouno et al., 'Making an orchestra speak', in *Sound and music computing* (2009), accessed 25 February 2026, <https://inria.hal.science/hal-00839067>.

77. Malherbe, *Locus*.

78. Niels Bogaards and Axel Röbel, 'An interface for analysis-driven sound processing', in *AES 119th Convention* (2005), 1–1, accessed 25 February 2026, <https://hal.science/hal-01161331v1>.

79. Jean Bresson, Carlos Agon and Gérard Assayag, 'OpenMusic: visual programming environment for music composition, analysis and research', in *Proceedings of the 19th ACM international conference on Multimedia* (2011), 743–746, <https://doi.org/10.1145/2072298.2072434>.

software to reconstruct the sources as a series of chord sequences to produce “delay” or “pedal” effects.⁸⁰

- Peter Ablinger’s *Voices and Piano* works⁸¹ for piano and tape take the recorded speech of several artists and philosophers. These works use the quadraturen method, a grid-based spectral analysis method inspired by photographic techniques, using a computer program developed by the composer (and Thomas Musil) with a particular focus on window size. These present a transparent comparison between speech and music in which ‘the piano part is the analysis of the voice.’ This method is also found in *Quadraturen III* (2009) for computer-controlled piano.
- Daniel Formo’s *The Orchestra of Speech*⁸² presents a software instrument that involves pre-prepared and live resynthesis of speech, realising resynthesis using transducers mounted on instruments including sitar, drums, and piano. The system also involved sample playback organised according to speech genre, including baby talk, small talk, and pillow talk. A Markov model was used to describe the likelihood of transitions between segments and, in reacting to live audio input, could trigger similar samples, responses, and alternative sequences.

Text-based or Visual-Based Mapping

These works explore mappings of text, formal grammar, and visual material. Of the seven examples, two could be described as generative systems, which are procedures for generating material rather than compositions themselves. Five used highly formalised systems in the mapping procedure, with two applying looser interpretations. In contrast to previous categories, most did not aim to produce speech-like sounds and instead took something of the structures of speech and text to produce new material with its own qualities.

- Vassallo’s *Versificator*⁸³ is an algorithmic music and text generation system that creates text and phonetic cues and uses these to generate musical material. Developed in Max-MSP, the system has three generator modules: a first module generates consonant strings,

80. Claudy Malherbe, ‘Locus: Rien n’aura eu lieu que le lieu’, in *The OM Composer’s Book 2*, ed. Jean Bresson, Carlos Agon and Gérard Assayag (DelaTour/Ircam, 2008).

81. Ablinger, *Voices and Piano*.

82. Formo, ‘The Orchestra of Speech—a speech-based instrument system’.

83. Vassallo, *Versificator*.

a second module generates vowel strings, and the last module generates ‘pseudowords’. Pitches are generated based on the translation of vowel formants to musical instruments, creating harmonic fields with durations scaled according to the vowel’s typical duration. The system is described as an ‘extension for the basic creative mechanism.’

2.4.4 Speech Synthesis

This section looks specifically at the use of synthesised speech in musical works. These are divided according to the speech synthesis method.

Linear Predictive Coding

This method was widely used in works by Paul Lansky (Six Fantasies, 1978), outside the scope of this review.⁸⁴

- DeMarinis’ *Alien Voices*⁸⁵ was an installation piece using a telephone booth, in which participants spoke into the receiver and heard the melody of their voices translated into music, whispers, robotic monotones, or ‘alien melodies’. The aim of the work was to reveal the pitch fluctuations of the audience’s voice through musical phrases.

Markov Models

- In Astrinaki’s *MAGE 2.0*⁸⁶ a guitar was used to control a speech synthesis system in real time. The text was a synthesised reading of Bishop’s poem The Fish, with guitar material divided into two gesture classes: continuous (bends, slides, etc.) and triggering (pitch, note on, note off, etc.). These gesture classes were used to control parameters including changing between voices, controlling speaking style, and manipulating vocal characteristics such as vocal tract length.

Neural Synthesis

- Jennifer Walshe and Memo Akten’s *ULTRACHUNK*⁸⁷ for voice and computer took as a source Walshe’s solo improvisations recorded on video over a year. An AI system was

84. David Loberg Code, ‘Observations in the Art of Speech: Paul Lansky’s Six Fantasies’, *Perspectives of New Music*, 1990, 144–169, <https://doi.org/10.2307/833348>.

85. DeMarinis, *Alien Voices*.

86. Astrinaki et al., ‘MAGE 2.0: New Features and its Application in the Development of a Talking Guitar’.

87. Akten and Walshe, *ULTRACHUNK*.

trained on Walshe’s voice and face (GRANNMA: Granular Neural Music and Audio), with video and audio generated live in duet with Walshe, creating an ‘improvisational duet between a classically-trained musician and her AI doppelganger’.⁸⁸

2.5 Discussion

2.5.1 Computational Methods

A summary of computational methods is shown in Table 2.2. Works were counted more than once if multiple methods were used.

Table 2.2: Computational methods, number of works in each category.

Method	Subcategory	Works
Sample playback	<i>Fixed tape tracks</i>	18
	<i>Live triggering</i>	5
Electronic manipulation	<i>Filtering</i>	18
	<i>Delays, reverb</i>	13
	<i>Loops and repetition</i>	10
	<i>Granulation</i>	1
	<i>Interpolation</i>	1
	<i>Combination</i>	2
	<i>Altering speed/direction</i>	10
	<i>Pitch shifting</i>	12
Mapping	<i>Instrumental resynthesis (Including custom made instruments, with or without electronics)</i>	14
	<i>Text-based or visual-based mappings</i>	3
Speech synthesis	<i>Linear Predictive Coding (LPC)</i>	5
	<i>Machine learning</i>	3

Some methods were overrepresented: sample playback was present in several works, but the majority were not defined by this method. Methods may be underrepresented if either the documentation did not mention them or if the results could not be heard. Accepting these limitations, the results showed a high proportion of signal processing methods including filtering,

88. Mark Dyer, Zubin Kanga and Jennifer Walshe, ‘In Conversation with Jennifer Walshe: Performing with Intelligent Machines’, *Contemporary Music Review* 42, no. 3 (2023): 391–399, <https://doi.org/10.1080/07494467.2023.2276563>.

delays, reverb, and pitch shifting. Instrumental resynthesis was also a common technique, from notated music based on spectral analysis to live filtering of instrumental sound.

The lack of machine-learning methods was surprising. This could be due to flaws in the methodology, lack of documentation, or a genuine lack of works. Two machine-learning methods used Markov chains. One work used neural synthesis, suggesting a space for new works using this method.

2.5.2 Sound Sources

Table 2.3 counts sound sources for works. If a work used sources from more than one category, it was counted twice.

Table 2.3: Number of works by sound source.

Sound source	Number of works
Vocal formants (non-specific speaker)	6
Casual conversation and interviews	7
Field recordings	2
Public figures	7
Spoken literature or poetry	8
Archival	8
Performer’s speech	6
Audience’s speech	2
Affected speech	4
Speech corpus	2
Text	3
Grammar structures	2
Synthesised speech	5

The largest categories were spoken literature, poetry, and archival sources, followed by casual conversation, interviews, and public figures. The prevalence of these categories suggests that speech was often used not just for its sonic qualities, but for its literary, historical, geographical, and political significance.

Performer’s speech, vocal formants defined as the resonant frequencies that determine the

quality of a speech sound, affected speech with deliberately exaggerated emotional or timbral qualities, and synthesised speech were represented with a middling number of examples. These sources allow flexibility through the deliberate shaping of the source to fit an artistic purpose. In many of these works, the focus is on the sonic qualities of speech, with the use of formant structures as a strategy to avoid contextual references associated with texts or speakers.

Working with an audience’s speech, a corpus, or formal grammar all involve unique technical challenges and may explain their underrepresentation in this study. However, they do suggest space for new works exploring these sources.

2.5.3 Artistic Intent

While the intent of the works differed, a small group of extra-musical themes emerged. These are summarised in Table 2.4; as before, works were repeated if multiple themes were identified.

Table 2.4: Number of works by shared theme.

Theme	Number of works
Human-machine interaction	12
Deliberately obscuring semantic meaning	12
Reproducing speech in a new medium	8
Highlighting geographical qualities of speech	4
Making a political or social statement	5

Works were not included if themes were particularly specific or opaque. The use of speech for its musical qualities could apply to almost every work, and so was not included. Likewise, the recorded speech in this study was naturally displaced from the body of the source or performer.

Human-machine interaction was a common theme. In some cases, this applied to preparatory stages of composition such as the composer’s interaction with automated tools, including the deliberate use of errors. In others, this theme was explored in performance with live electronics and sound-making machines. In the works identified, this was a positive or neutral relationship.

Deliberately obscuring semantic meaning was widely explored, often involving degrees of semantic clarity as part of a compositional framework. However, this theme was often unavoidable as extensive signal processing would always result in a loss of semantic meaning.

Eight works reproduced speech in a new medium, from orchestral instruments to player piano. The process of reproduction was a central theme, as the materials for the reproduction were repurposed to that end. In several works, the success of the method was measured by the intelligibility of the reproduction, and the successful results in this respect were Ablinger's *Voices and Piano*, Hummel's *Nicanor*, and Harvey's *Speakings*. Riches's *Talking Machine* also produced intelligible speech but was designed as an acoustic speech synthesiser.

Highlighting the geographical qualities of speech was not particularly common, but four works used recordings of the same phrase, word, or phoneme in multiple languages.⁸⁹

Political and social statements did feature, but not widely. These works added significance to the use of speech through the speaker's identity, associations, and the content of the speech. These included the war in Ukraine, the speech patterns of dictators, and positive statements on the unifying qualities of language.

2.5.4 Instrumentation

Table 2.5 looks at the medium of works in live performance. Works were counted twice if they used both tape and live electronics.

Table 2.5: Number of works by instrumentation.

Instrumentation	Instruments Alone	With Tape	With Live Electronics
Tape or video works	–	18	–
Digital musical instruments	1	–	–
Solo instrument or voice	1	2	6
Ensemble	3	8	4
Orchestra	3	1	1
Mechanical devices	3	0	0
Other instruments	3	0	1

Most works were solely for tape, and a large proportion of instrumental works were augmented using tape or live electronics. The *other instruments* category included turntables, a

89. Wishart, *Globalalia*; Blablabor, *hirsch hirn hornisse*; Gál, *bestimmung new york*; Bosetti, *Alpine Flipbooks*.

vibrating handheld device, and a telephone booth.

There were three generative systems identified. These compositional tools were without a specific artistic output and so were not included in the table.

2.5.5 Compositional Challenges

Ciardi outlines challenges in speech transcription for musical instruments, including playability and making a speech reference evident despite ambiguities caused by the transcription process.⁹⁰ Visual information (often text projections) is identified as a strategy to increase the intensity of a ‘prosodic reference’; to elevate the listening experience above the vague perception that instruments are imitating speech. This was not observed in this study and instead, a common technique involved presenting speech audio or spoken text alongside instrumental resynthesis.

The use of speech can raise ethical questions: James Tenney’s *Ain’t I a Woman?*⁹¹ used Sojourner Truth’s speech on anti-slavery resistance and women’s suffrage. A decision was made not to program this piece in Toronto, 2017 after consultation with the community, feeling that it was ‘inappropriate to program a work of white art based on black pain’,⁹² and prompting a discussion around the appropriateness of the artistic use of certain kinds of speech. The composer Juliana Hodkinson has described the importance of an ethically informed artistic position when speech refers to ‘difficult histories’.⁹³

2.5.6 Trends

A tendency toward live methods was observed, including live triggering, live spectral analysis and resynthesis, and the live use of Markov Models and neural synthesis in synthesised speech. More typical existing methods such as looping, delays, and reverb were found to be increasingly applied in a live setting, often in improvisational works.

Automation of compositional processes was found in recent works, particularly regarding

90. Fabio Cifariello Ciardi, ‘Strategies and tools for the sonification of prosodic data: A composer’s perspective’, in *Proceedings of the 26th International Conference on Auditory Display (ICAD 2021)* (2021), 247–252, <https://doi.org/10.21785/icad2021.041>.

91. James Tenney, *Ain’t I a Woman?*, 1992, accessed 25 February 2026, <https://brahms.ircam.fr/en/works/work/46265/>.

92. W. Bartley, *Concert Report: The Fine Art of Social Consciousness*, 2017, accessed 25 February 2026, <https://thewholenote.com/index.php/newsroom/blog/concert-reviews/27382-concert-report-the-fine-art-of-social-consciousness-james-tenney-resistance>.

93. Juliana Hodkinson, ‘Resounding difficult histories’, in *Over and Over and Over Again: Reenactment Strategies in Contemporary Arts and Theory*, ed. Cristina Baldacci, Clio Nicastro and Arianna Sforzini (ICI Berlin Press, 2022), 69–80.

the transcription and orchestration of speech, and the creative use of errors resulting from these processes.

One further trend involved the use of acoustic-mechanical instruments, identified in five of the works (not including player piano works). With the proliferation of microcontrollers, more work in this area is expected.

2.5.7 Previous Research

There were two field reviews closely related to this subject. Lane's was the most comprehensive,⁹⁴ dividing works according to compositional method and intent, with categories based on the source (archival, conversational, spoken word). Subgroups focused on the transmission of semantic meaning, including dissolution or accumulation through preprocessing, deconstruction, translation, and more. Lane suggests a division between composers who use speech for its real-world associations, and the acousmatic tradition of electroacoustic music. The results of this study support this, in the difference between works using speech for its political, social, geographical, and semantic content, and those interested primarily in the sonic qualities of speech.

A systematic review of NIME publications relating to the voice used automated natural language processing as a step in thematic analysis.⁹⁵ A taxonomy based around these categorical themes looked at the intention of the research, the application, and finally the implementation. The review describes voice as a 'bridge between diverse areas of research and artistic practice', a finding reflected in this review.

Segini and Ruviano present a broad overview of fifteen works involving music and speech from 1912.⁹⁶ The authors define two dimensions for analysis: the degree of text intelligibility, and position on a perceptual scale of music-speech, plotting the trajectory of selected works through this dimensional space. The use of this space is suggested as a compositional tool.

Bossis identifies challenges in musicological analysis of electroacoustic works dealing with 'artificial vocality,' suggesting that traditional models are redundant.⁹⁷ To address an 'absence

94. Lane, 'Voices from the Past: Compositional approaches to using recorded speech'.

95. Kleinberger et al., 'Voice at NIME: a Taxonomy of New Interfaces for Vocal Musical Expression'.

96. Rodrigo Segnini and Bruno Ruviano, 'Analysis of Electroacoustic Works with Music and Language Intersections', in *ICMC 2005* (2005), 5v, accessed 25 February 2026, <http://hdl.handle.net/2027/spo.bbp2372.2005.156>.

97. Bruno Bossis, 'Reflections on the analysis of artificial vocality: representations, tools and perspective', *Organised Sound* 9, no. 1 (2004): 91–98, <https://doi.org/10.1017/S1355771804000123>.

of pitches ... [and] rarity of rhythms with a regular pulse', continuous measures of sound qualities are proposed including a 'vocality index' describing the proximity of a sound to voice, and the pulsed character of the music.

Results outside the scope of this review included Anhalt's *Orpheus resurgent ... perhaps* which looks at voice from an acoustic, psychological, and compositional perspective, structured around vocal techniques, phonetic constructs, vocal tract behaviour, and compositional treatment of text.⁹⁸

This review builds on Lane's work, but groups examples according to compositional method rather than source material or transmission of meaning. As a result of the festival program searches, several works in this review were not identified in previous research.

2.5.8 Limitations

This study was limited by the sources; electronic music festivals or record releases were outside the scope of this study but would return relevant results. Data not captured included work around neural synthesis, including RAVE models.⁹⁹ While these have scope beyond speech (a likely reason for the omission), several models were developed from speech datasets.

Relevant work was likely missed through keyword exclusion; several words used as metaphors were also key search terms ('language' or 'voice'). A similar challenge arose in avoiding research on speech and music discrimination. Several practice-based examples did not have sufficient documentation to be included. An effort was made to include as many works as possible, so where works had only a small amount of information they were included.

2.6 Conclusion

This field review examined a body of research and practice using technology to create and perform speech-derived musical works, highlighting the innovative and diverse practices in this area.

Results showed widespread use of signal processing techniques and novel methods of instrumental resynthesis. Speech synthesis and generative systems have found unique applications in

98. Istvan Anhalt, 'Alternative Voices: Essays on contemporary vocal and choral composition', in *Alternative Voices* (University of Toronto Press, 1984), <https://doi.org/10.3138/9781442656314>.

99. Antoine Caillon and Philippe Esling, 'RAVE: A variational autoencoder for fast and high-quality neural audio synthesis', *CoRR*, 2021, <https://doi.org/10.48550/arXiv.2111.05011>.

several works, underlining common themes including human-machine interaction, the deliberate obscuration of semantic meaning, the reproduction of speech in new mediums, and the use of speech for making geographical, political, or social statements. Most works used tape, with instrumental works often supplemented with live electronics or acoustic-mechanical instruments, showing a growing trend towards live processing and resynthesis. A trend toward automated compositional methods was also observed.

Chapters 4, 5, and 6 are each informed by the findings of this field review. In Chapter 4 I draw on approaches to instrumental resynthesis, automated transcription and live improvisation, with *Motorcase II* taking the relatively underexplored speech corpora as a compositional source. In the same chapter, methods of electronic manipulation described in this review are applied to the control of electromechanical devices, through the interfaces of *Motorcase I* and *II*. In Chapter 5, I pair machine-learning-based speech synthesis with electromechanical sound-production through *Smalltalk* and *Radio User Guide*. The use of grammar structures, also identified here as an underexplored source, become the source material for the creative work in Chapter 6, through *Porcelain Music Box*, *Seven Studies for Electric Motors*, and *Word games*.

Chapter 3

Machine Learning in Electromechanical Sound Art

3.1 Introduction

While the previous chapter surveyed computational methods in speech-derived music, the tools available for this practice have been fundamentally expanded by the recent integration of machine learning. As I later employ RAVE¹ and text models in the development of new creative work (Chapter 5, Chapter 6), in this chapter I examine the existing applications of neural networks within electromechanical sound art. Because works combining electromechanical elements with speech-derived musical material remain relatively rare, this review has a broader scope—including, but not limited to those which use speech and language as a data source. This allows for a focus on the distinct affordances of physical sound production, providing necessary context for the creative works in the submitted portfolio.

Neural networks have found wide adoption in a range of musical contexts over the past decade, primarily focused on the symbolic and digital domains,² but there are comparatively few documented examples of their use within electromechanical sound producing systems. Recent works illustrate this practice, from Gardener and Debicki’s Empty Vessels (2019) in which

1. Antoine Caillon and Philippe Esling, ‘RAVE: A variational autoencoder for fast and high-quality neural audio synthesis’, *CoRR*, 2021, <https://doi.org/10.48550/arXiv.2111.05011>.

2. Jean-Pierre Briot, ‘From artificial neural networks to deep learning for music generation: history, concepts and trends’, *Neural Computing and Applications* 33, no. 1 (2021): 39–65, <https://doi.org/10.1007/s00521-020-05399-0>; Lei Wang et al., ‘A review of intelligent music generation systems’, *Neural Computing and Applications* 36, no. 12 (2024): 6381–6401, <https://doi.org/10.1007/s00521-024-09418-2>.

mechatronic cellos perform generated MIDI sequences, to Quadrature's *Fantasie#4* (2019) which used radio wave data as a seed for generative models and player piano. To document artistic approaches in this emerging field, this chapter presents a practice review of works from the past decade, identifying how physical constraints shape the design of these systems and discussing the varied forms of interaction that arise.

This review was guided by three questions:

1. How are neural networks currently used in electromechanical music and sound art, and which areas remain underexplored?
2. How do material constraints shape the design and function of these systems?
3. What approaches to interaction are employed in these systems, and how can they inform new modes of creative practice?

Related work in this field includes Parkinson and Dunning's taxonomy of improvisational interaction with musical machines,³ and broader surveys of robotic musicianship.⁴ Several frameworks have emphasised the artistic dimensions of combining computational and material processes, including Valle's description of acoustic computer music in which the limitations of found objects and actuators are seen as a creative affordance,⁵ and Pigott's emphasis on the electromechanical mechanism itself as central to creative exploration.⁶ Zareei's *Audiovisual Materialism* argues that the visual and material aspects of this practice are inseparable from the sonic,⁷ connecting with a broader field of practice including kinetic sound art and sculpture. The following section outlines this context, tracing a renewed interest in mechanically-produced sound and its relevance to current forms of machine learning-based musical systems.

3. Adam Parkinson and Graham Dunning, 'Improvising with Machines: A taxonomy of musical interactions', *Organised Sound* 29, no. 3 (2024): 340–351, <https://doi.org/10.1017/s1355771824000268>.

4. Mason Bretan and Gil Weinberg, 'A survey of robotic musicianship', *Communications of the ACM* 59, no. 5 (2016): 100–109, <https://doi.org/10.1145/2818994>.

5. Andrea Valle, 'Making acoustic computer music: The Rumentarium project', *Organised Sound* 18, no. 3 (2013): 242–254, <https://doi.org/10.1017/s1355771813000216>.

6. Jon Pigott, 'Across Fields: Sound, art and technology from an electromechanical perspective', *Organised Sound* 22, no. 2 (2017): 276–285, <https://doi.org/10.1017/s1355771817000188>.

7. Mo H Zareei, 'Audiovisual materialism', *Organised Sound* 25, no. 3 (2020): 362–371, <https://doi.org/10.1017/s1355771820000321>.

3.2 Electromechanical Music

Current practitioners are drawing on a history of kinetic sound art and mechanical music, with automatic instruments paired with algorithmic procedures as early as Winkel's 1821 *Componium*, a self-playing organ capable of generating variations on a theme.⁸ The late twentieth century resurgence of interest in mechanical instruments saw the expressive potential of the player piano expanded notably by Nancarrow and later by Peter Ablinger and others.⁹ Mechatronic ensembles including LEMUR, Machine Orchestra, and SolidNoise Ensemble represent a growing interest in the expressive potential of automated instruments, computational control, and DIY methods, alongside a recognition of the unique character and qualities of acoustic sound.¹⁰

At the same time, kinetic sculpture and electromechanical sound art have been developing throughout the late twentieth and early twenty-first century, with notable examples including Trimpin's interactive sculptures and installations.¹¹ Other artists working in this field include Peter Vogel, Pe Lang, Zimoun, and Marianthi Papalexandri, among many others.¹²

Despite the widespread use of algorithmic control in electromechanical musical systems, the incorporation of machine learning into this field is a relatively recent practice, with exceptions including Baginsky's self-learning robotic rock band The Three Sirens (1992-).¹³ Unlike purely electronic systems, electromechanical systems introduce a materiality that shapes both the sound produced and the conceptual approach taken by the artist. These systems present a particularly compelling case for exploring machine learning algorithms through their acoustic and material properties, requiring consideration of the instruments' physical constraints and sonic affordances,

8. Heinz Zemanek, 'Geschichte des Automaten: Das Componium von Winkel Ein Gerät mit einem Zufallsgenerator aus dem Jahre 1821', *it-Information Technology* 8, nos. 1-6 (1966): 61-62, <https://doi.org/10.1524/itit.1966.8.16.61>.

9. Peter Ablinger, *Quadraturen III*, 2004, accessed 25 February 2026, <https://ablinger.mur.at/docu11.html#qu3>; Helena Bugallo, 'Selected "Studies" for player piano by Conlon Nancarrow: Sources, working methods, and compositional strategies' (PhD diss., State University of New York at Buffalo, 2004); Phillip Venables, *Answer Machine Tape, 1987*, 2022, accessed 25 February 2026, <https://philipvenables.com/2022/09/25/answer-machine-tape-1987>; Asbjørn Blokkum Flø, 'Materiality in sound art', *Organised Sound* 23, no. 3 (2018): 225-234, <https://doi.org/10.1017/s1355771818000134>.

10. Jiffer Harriman et al., 'SolidNoise: Making Musical Robots', in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (2016), 2504-2510, <https://doi.org/10.1145/2851581.2892539>; Ajay Kapur et al., 'The machine orchestra: An ensemble of human laptop performers and robotic musical instruments', *Computer Music Journal* 35, no. 4 (2011): 49-63, https://doi.org/10.1162/comj_a_00090.

11. Sasha Leitman, 'Trimpin: an interview', *Computer Music Journal* 35, no. 4 (2011): 12-27, https://doi.org/10.1162/comj_a_00088.

12. Pigott, 'Across Fields: Sound, art and technology from an electromechanical perspective'.

13. Sofian Audry, Victor Drouin-Trempe and Ola Siebert, 'The Strangest Music in the World: Self-Supervised Creativity and Nostalgia for the Future in Robotic Rock Band "The Three Sirens"', in *Arts*, vol. 12, 1 (MDPI, 2022), 2, <https://doi.org/10.3390/arts12010002>.

from actuation mechanisms to the resonant properties of materials. At the same time, machine learning applications can extend approaches to interaction by allowing features to be learned from data rather than being explicitly programmed, as well as facilitating novel generative behaviours. To understand how artists are working within this space, the following section outlines the methodology used to survey existing works and structure a review of current approaches.

3.3 Methodology

First, a review of existing works from 2015 to the present was undertaken to provide a survey of the field over the past decade. The search began with academic databases,¹⁴ journals,¹⁵ and conference proceedings.¹⁶ Terms covered three areas: machine learning, mechatronic music, and sound art, combining at least one machine learning term (such as ‘AI’ or ‘neural network’) with at least one mechatronic term (such as ‘robotic’, ‘solenoid’ or ‘kinetic’) and one sound art or music term (such as ‘sound art’ or ‘installation’). To reduce the results to a manageable size, limitations were applied to the subject area or relevant keywords. This approach highlighted the limitations of keyword search strategies, e.g., kinetic is a common musical metaphor, which can obscure results. Results were supplemented with a search of gallery archives, artist websites, and repositories to identify further artistic projects and performances not otherwise documented.¹⁷ This survey is partial, and reflects available documentation rather than an exhaustive catalogue.

3.4 Results

Surveyed works were manually organised into categories reflecting the function of a neural network. Many of the most prominent applications involve models that are highly specialised in their domain; MIDI generation models depend on the structure of symbolic music, and speech-to-text models are designed to process phonetic and linguistic data. In these cases, the categories are partly defined by the data itself (e.g., MIDI Generation, Speech Transcription). However, a

14. RILM, JSTOR, SCOPUS, IEEE Xplore

15. Organised Sound, Leonardo, Tempo, Neural, Journal of New Music Research, Computer Music Journal, Journal of Creativity

16. NIME, ICCM, ICMC, AIMC (keyword search terms)

17. such as Instrumentinventors.org, Ars Electronica, Research Catalogue, V2 Lab for the Unstable Media, Art Machines 2 Symposium (2021)

flexible tool like Wekinator¹⁸ is not defined by a specific domain and so the category Live Parameter Mapping is used to describe its role in creating real-time, learned connections between input and output.

3.4.1 Audio Classification

Software libraries for machine learning based audio classification are becoming increasingly accessible, including Flucoma¹⁹ and MLStar.²⁰ While neural networks have been applied to audio classification since the late 1990's, these recent tools allow practitioners without expert knowledge to train a neural network on a dataset of categorised sounds, using a set of user-defined audio features to teach the model to distinguish between different sound classes. These features can describe timbral characteristics, such as Mel-frequency cepstral coefficients (MFCCs), brightness, or roughness, or rhythmic qualities.

In practice, audio classification is often combined with acoustic instruments in performer-responsive works, with two works in this category exploring parallel human-robotic configurations.

- Ben Eyes' *Ghost in the Piano*²¹ augments a piano with motors, solenoids, actuators, and loudspeakers, analysing live guitar audio using FluCoMa²² to classify playing techniques, which then inform a wavetable-based sequencer and synthesiser. The resulting accompaniment is projected through the piano's acoustic resonance via synthesised sound and motor-driven actuators.
- Gioti's *Imitation Game*²³ paired a human percussionist with a robotic counterpart, con-

18. Rebecca Fiebrink and Perry R Cook, 'The Wekinator: a system for real-time, interactive machine learning in music', in *Proceedings of the eleventh international society for music information retrieval conference*, vol. 3 (International Society for Music Information Retrieval, 2010), 2–1, accessed 25 February 2026, <https://ismir2010.ismir.net/proceedings/late-breaking-demo-13.pdf>.

19. Pierre Alexandre Tremblay, Gerard Roma and Owen Green, 'Enabling programmatic data mining as musicking: The fluid corpus manipulation toolkit', *Computer Music Journal* 45, no. 2 (2021): 9–23, https://doi.org/10.1162/comj_a_00600.

20. Benjamin D Smith and W Scott Deal, 'ML.* Machine learning library as a musical partner in the computer-acoustic composition flight', in *Proceedings of the 2014 International Computer Music Conference* (Michigan Publishing, 2014), 1285–1289, accessed 25 February 2026, <https://hdl.handle.net/1805/15340>.

21. Gabriel Kopito et al., 'Ghost in the Piano – An Investigation into AI, Robotics and Musical Performance', in *AIMC 2024* (2024), <https://doi.org/10.5281/zenodo.15110248>.

22. Pierre Alexandre Tremblay et al., *The Fluid Corpus Manipulation Toolbox (v.1)*, v. v1.0.0, 2022, <https://doi.org/10.5281/zenodo.6834643>.

23. Artemi-Maria Gioti, 'Imitation game: Real-time decision-making in an inter-active composition for human and robotic percussionist', in *Proceedings of the International Computer Music Conference* (2019), accessed 25 February 2026, https://interagency.iem.at/pdfs/Gioti_ICMC2019.pdf.

trolled by a feed-forward neural network trained to classify the rhythms, instruments and playing techniques of the human performer in three modes of interaction: repeat, imitate, and initiate.

- Franklin’s *Robo-Cajon*²⁴ also features a parallel configuration, pairing a solenoid-actuated cajon with a human-played cajon equipped with a contact microphone, classifying rhythmic data from the human performer using a neural network in MLStar,²⁵ with a hidden Markov model (HMM) used to predict and generate rhythmic patterns.

3.4.2 MIDI Generation

Machine learning-based MIDI generation²⁶ typically involves training a neural network, such as a recurrent neural network²⁷ (RNN) or a transformer,²⁸ on a large dataset of existing MIDI. Once trained, the models can generate new sequences that reproduce the stylistic and structural features of the source material, though challenges remain in capturing long-term form and expression.²⁹

These applications are well suited to existing mechanical instruments such as the Disklavier, due to their compatibility with MIDI control and the availability of extensive training datasets,³⁰ while projects involving custom-built instruments often require the creation of new datasets that reflect the specific affordances of the instrument. However, both approaches often feature other statistical or rule-based methods in tandem with generative models.

- Quadrature’s *Fantasie#4*³¹ mapped live radio telescope signals to player piano, generating

24. Austin Franklin, ‘The Robo-Cajon: An Example of Live Performance with Musical Robotics’, in *International Computer Music Conference, ICMC Proceedings* (International Computer Music Association, 2024), 270–273, accessed 25 February 2026, <https://urn.kb.se/resolve?urn=urn:nbn:se:mdh:diva-70059>.

25. Smith and Deal, ‘ML.* Machine learning library as a musical partner in the computer-acoustic composition flight’.

26. Shulei Ji, Xinyu Yang and Jing Luo, ‘A survey on deep learning for symbolic music generation: Representations, algorithms, evaluations, and challenges’, *ACM Computing Surveys* 56, no. 1 (2023): 1–39, <https://doi.org/10.1145/3597493>.

27. Daniel Jurafsky and James H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*, 3rd (2026), 13, accessed 25 February 2026, <https://web.stanford.edu/~jurafsky/slp3/>.

28. Ashish Vaswani et al., ‘Attention is all you need’, *Advances in neural information processing systems* 30 (2017).

29. Rohan Mitra and Imran Zuolkernan, ‘Music generation using deep learning and generative AI: a systematic review’, *IEEE Access*, 2025, <https://doi.org/10.1109/access.2025.3531798>.

30. Wang et al., ‘A review of intelligent music generation systems’.

31. Quadrature, *Fantasie 4*, 2021, accessed 25 February 2026, <https://quadrature.co/work/fantasie4/>.

MIDI using Piano Genie,³² and Magenta’s Polyphony-RNN³³ trained on twentieth-century piano music. The work presents a gradual divergence of machine learning responses from direct mappings of the radio signals, as the models elaborate on the data.

- Nikrang’s *Absolute Hallucination*³⁴ used Ricercar,³⁵ a collaborative software trained on public domain classical works, to control a self-playing piano. The work explores random compositional paths through the ‘creative space’ of the model.
- Thompson IV³⁶ created duets for Disklavier and human performer, also training Polyphony-RNN on MIDI transcribed from speech, and piano music from Thelonious Monk to Mara Gibson.
- Refree’s *Interleggere ad affectum*³⁷ (2025) repurposed GrooveTransformer,³⁸ a transformer-based model for real-time drum accompaniment generation, to produce a rhythmic and harmonic accompaniment in a work for multiple pipe organs. The work used a Markov-based process to generate harmonic content and note durations, with a second performer supervising the generative accompaniment.
- Arca, Armitage and Shepardson’s *The Light Comes in the Name of the Voice*³⁹ employed a generative system built around the Notochord MIDI model⁴⁰ to extend Arca’s improvisations on the Magnetic Resonator Piano, an electromagnetically augmented grand piano,⁴¹ into ‘uncanny’ reinterpretations.

32. Chris Donahue, Ian Simon and Sander Dieleman, ‘Piano genie’, in *Proceedings of the 24th International Conference on Intelligent User Interfaces* (2019), 160–164, <https://doi.org/10.1145/3301275.3302288>.

33. Magenta Team, *Magenta: Music and Art Generation with Machine Intelligence*, 2016, accessed 25 February 2026, https://github.com/magenta/magenta/tree/main/magenta/models/polyphony_rnn.

34. A. Nikrang, *Absolute Hallucination*, 2024, accessed 25 February 2026, <https://ars.electronica.art/futurelab/en/projects-absolute-hallucination/>.

35. Ali Nikrang and Susanne Kiesenhofer, ‘Human–AI Co-Creation in Contemporary Composition: Interaction and Artistic Strategies with Ricercar’, in *Proceedings of the Conference on Animation and Interactive Art* (2025), 65–73, <https://doi.org/10.1145/3749893.3749961>.

36. W. A. Thompson IV, ‘Creating Musical Scores Inspired by the Intersection of Human Speech and Music Through Model-Based Cross Synthesis’ (PhD diss., Louisiana State University, 2022).

37. Nicholas Evans, Behzad Haki and Sergi Jordà, ‘Repurposing a Rhythm Accompaniment System for Pipe Organ Performance’, in *Proceedings of the International Conference on New Interfaces for Musical Expression* (2025), 116–120, <https://doi.org/10.5281/zenodo.15698807>.

38. Behzad Haki, Nicholas Evans and Sergi Jordà, ‘GrooveTransformer: A Generative Drum Sequencer Eurorack Module’, in *Proceedings of the International Conference on New Interfaces for Musical Expression* (2024), 261–265, <https://doi.org/10.5281/zenodo.13904848>.

39. Jack Armitage and Victor Shepardson, ‘Augmenting the expressivity of the Notochord generative midi model for Arca’s “The Light Comes in the Name of the Voice”’, in *Proceedings of the 5th Conference on AI Music Creativity* (2024), accessed 25 February 2026, <https://aimc2024.pubpub.org/pub/0lh6s86c>.

40. Victor Shepardson, Jack Armitage and Thor Magnusson, ‘Notochord: a flexible probabilistic model for real-time MIDI performance’, *arXiv preprint arXiv:2403.12000*, 2024, <https://doi.org/10.48550/arXiv.2403.12000>.

41. Andrew McPherson, ‘The magnetic resonator piano: Electronic augmentation of an acoustic grand piano’, *Journal of New Music Research* 39, no. 3 (2010): 189–202, <https://doi.org/10.1080/09298211003695587>.

- Gardener and Debicki's *Empty Vessels*⁴² features a mechatronic cello trio performing generated MIDI sequences, using an HMM to generate melody parts from the MIDI model's harmonic output.
- Benjamin Bacon's *Little Sound Machines*⁴³ consists of multiple devices constructed from salvaged and repurposed objects. The work used separate Tensorflow⁴⁴ models for rhythm, harmony and melody construction, followed by stochastic parameter generation in Max-MSP (Cycling '74). Throughout the installation, the models learn from, and influence each other in a 'purely machine-made performance'.

3.4.3 Live Parameter Mapping

Mapping non-musical data streams to control mechanical actions is a common strategy in electromechanical sound art. In works not using machine learning, this is often achieved using rule-based systems, from Zareei's mapping of cellular automata to motor frequencies and LED patterns in Noise Square,⁴⁵ to the mapping of a dancer's movement to mechanically-actuated percussion in Hsu and Kemper's Why Should Our Bodies End at the Skin?,⁴⁶ and Eacott's submerged sensor in Floodtide⁴⁷ for translating tidal data into parameters for a mechanical organ.

The advantage of machine learning in this context is the ease with which complex and high-dimensional mappings can be created, with the relationships learned from examples provided by the user. Wekinator⁴⁸ is a prominent tool for real-time model training, supporting a range of algorithms – from neural networks to decision trees and support vector machines – and has

42. G. Debicki and D. Gardener, *Empty Vessels*, 2019, accessed 25 February 2026, <https://ars.electronica.art/newdigitaldeal/en/empty-vessels/>.

43. Benjamin Bacon, *Little Sound Machines*, 2018, accessed 25 February 2026, <https://www.benjaminbacon.studio/little-sound-machines>.

44. Martín Abadi et al., 'TensorFlow: A system for Large-Scale machine learning', in *12th USENIX symposium on operating systems design and implementation (OSDI 16)* (2016), 265–283, <https://doi.org/10.48550/arXiv.1605.08695>.

45. Mo H Zareei, Dale A Carnegie and Ajay Kapur, 'Noise Square: Physical Sonification of Cellular Automata through Mech-atronic Sound-sculpture', in *Proceedings of the 21st International Symposium on Electronic Art* (Vancouver, Canada, 2015), accessed 25 February 2026, https://www.isea-archives.org/docs/2015/proceedings/ISEA2015_proceedings.pdf.

46. Aurie Hsu and Steven Kemper, 'Enacting Sonic-Cyborg Performance through the Hybrid Body in Teka-Mori and Why Should Our Bodies End at the Skin?', *Leonardo Music Journal* 29 (2019): 83–87, https://doi.org/10.1162/lmj_a_01069.

47. John Eacott, 'Flood Tide: sonification as musical performance—an audience perspective', *AI & society* 27 (2012): 189–195, <https://doi.org/10.1007/s00146-011-0338-2>.

48. Fiebrink and Cook, 'The Wekinator: a system for real-time, interactive machine learning in music'.

found wide applications in interactive music,⁴⁹ as well as the single example identified below in the electromechanical domain.

- Macionis and Kapur’s *Where is the Quiet*⁵⁰ used Wekinator⁵¹ to analyse EEG data (Muse headband), mapping alpha waves to solenoid-actuated singing bowls in response to a calm state of the participant.

3.4.4 System Modelling

System modelling involves training a model to reproduce the dynamics of a complex system. Unlike live parameter mapping, where an input is directly mapped to an output, this method uses a neural network to generate data that simulates or predicts a system’s behaviour. The underlying process is often similar to MIDI generation, in that a model learns sequential patterns from a dataset and is then used to make predictions – but the prevalence and specific focus of MIDI generation warrant a separate category.

- In *Supraorganism*,⁵² Justine Emard trained a neural network on data from a honeybee colony. The network’s predictions of the colony’s behaviour are embodied by robotic glass and light sculptures that produce sound by rotating wires against their surfaces.

3.4.5 Analogue Neural Networks

Some artists have used analogue models of neural networks as the basis for physical installation works. This approach can be seen in analogue electronic works such as Wolfgang Spahn’s *Echo State Network*⁵³ and with sound enacted through motors in Faubel’s generation of emergent rhythmic patterns using ‘neurally inspired’ analogue oscillators.⁵⁴

49. Rebecca Fiebrink and Laetitia Sonami, ‘Reflections on eight years of instrument creation with machine learning’, in *Proceedings of the International Conference on New Interfaces for Musical Expression* (2020), 237–242, accessed 25 February 2026, <https://research.gold.ac.uk/id/eprint/28405>.

50. McLean J Macionis and Ajay Kapur, ‘Where is the quiet: Immersive experience design using the brain, mechatronics, and machine learning’, in *Proceedings of the International Conference on New Interfaces for Musical Expression* (2019), 335–338, <https://doi.org/10.5281/zenodo.3672984>.

51. Fiebrink and Cook, ‘The Wekinator: a system for real-time, interactive machine learning in music’.

52. Justine Emard, *Supraorganism*, 2020, accessed 25 February 2026, <https://justineemard.com/supraorganism/>.

53. Wolfgang Spahn, *Echo State Network*, 2019, accessed 25 February 2026, http://wolfgang-spahn.com/doku.php/installation:echo_state_networks.

54. Christian Faubel, ‘Emergent Polyrhythmic Patterns with a Neuromorph Electronic Network’, in *Proceedings of the International Conference on New Interfaces for Musical Expression* (2021), <https://doi.org/10.21428/92fbeb44.e66a8542>.

- Spahn and Faubel’s *Symbolic Grounding*⁵⁵ is an installation work in which kinetic sound and light patterns emerge from an analogue network of artificial neurons – a process the artists describe as rendering the emergent ‘temporal structures and functions tangible’. The work’s use of neural oscillators makes this a distinct approach to the algorithmic machine learning strategies discussed elsewhere.

3.4.6 Speech Transcription

The transcription of speech to generate musical material has been widely used in the broader electromechanical field. Many works, from Peter Ablinger’s *Quadraturen*⁵⁶ series to Canuto’s stepper motor-driven *Talking Motors*⁵⁷ focus on spectral features, mapping the acoustic properties of the voice to automatic instruments or motors. Other approaches include physical models of the acoustic vocal tract, as found in the resonators of Ali Miharbi’s *Whispering*⁵⁸ series.

A different approach has emerged with the development of speech-to-text systems. These function by using an acoustic model to analyse an audio signal and map its sound features to phonemes. A language model then processes this sequence, statistically determining the most probable series of words. Artistic works using text transcription often centre around the misinterpretation or fragmentation of meaning.

- Benjamin Bacon’s *PROBE II, Subaudition*⁵⁹ used DeepSpeech to transcribe audience speech, transmitted to a second machine which translates the data to a mirrored zoetrope projection. The work focuses on the ‘translation, degradation, and misinterpretation of meaning’.

3.4.7 Text and Language Generation

Modern neural network models for text generation use an encoder-decoder framework to map an input sequence into low-dimensional vectors called embeddings. A decoder then uses these

55. C. Faubel and W. Spahn, *Symbolic Grounding*, 2018, accessed 25 February 2026, <https://christian.faubel.derstrudel.org/symbolic-grounding/>.

56. Peter Ablinger, *Quadraturen*, 1995 -, accessed 25 February 2026, <https://ablinger.mur.at/docu11.html>.

57. Jesus Canuto, *Talking Motors*, 2023, accessed 25 February 2026, <https://www.jesuscanuto.com/talking-motors>.

58. C. Faubel and W. Spahn, *Whispering*, 2016-2022, accessed 25 February 2026, <https://www.alimiharbi.com/work/whisper/>.

59. Benjamin Bacon, *PROBE II, Subaudition*, 2021, accessed 25 February 2026, <https://www.benjaminbacon.studio/probe-1>.

embeddings to generate new text.⁶⁰

Text generation has found artistic uses across music and sound art to generate musical forms, from Walshe’s *Text Score Dataset*⁶¹ to the deliberate undertraining of models in Kiratlı’s *Cacophonic Choir*.⁶² In the material domain, these models can be used to imbue physical objects with personality and narrative, and model non-human forms of communication.

- Jan Zuiderveld’s *Conversations beyond the ordinary*⁶³ embedded large language models in appliances—a coffee machine, a microwave, and a photocopier—giving the devices specific personalities; the coffee machine requires the audience’s sincere desire for a coffee in order to produce one.
- Savery developed singing capabilities for the robotic marimba player, *Shimon*,⁶⁴ generating lyrics with a combination of WordNet to provide synonyms and antonyms from user-defined keywords, an LSTM for text generation, and a Markov model to refine the generated rhymes. The authors synchronise the robot’s physical gestures with the lyrics, including ‘emotion-conveying eyebrow movements’.
- Thomas Grill’s *Mutual Understanding* installation⁶⁵ uses two agents communicating via horn loudspeakers, inspired by an experiment in which chat bots were directed to optimise their language for negotiation. The agents develop a vocabulary independent from human language with a goal to ‘maximise the beauty of their own vocal expression’.

3.5 Discussion

The following discussion considers the guiding questions of this review: identifying current and emerging applications, how material constraints shape the design and function of electromech-

60. Junyi Li et al., ‘Pre-trained language models for text generation: A survey’, *ACM Computing Surveys* 56, no. 9 (2024): 1–39, <https://doi.org/10.1145/3649449>.

61. Jennifer Walshe, *The Text Score Dataset 1.0*, 2021, accessed 25 February 2026, <http://milker.org/s/The-Text-Score-Dataset-10-Jennifer-Walshe.pdf>.

62. Şöl en Kiratlı, Hannah E Wolfe and Alex Bundy, ‘Cacophonic choir: an interactive art installation embodying the voices of sexual assault survivors’, in *ACM SIGGRAPH 2020 Art Gallery* (Association for Computing Machinery, 2020), 446–450, <https://doi.org/10.1145/3386567>.

63. J. Zuiderveld, *Conversations beyond the ordinary*, 2024, accessed 25 February 2026, <https://ars.electronica.art/hope/en/conversations-beyond-the-ordinary/>.

64. Richard Savery, Lisa Zahray and Gil Weinberg, ‘Shimon sings-robotic musicianship finds its voice’, in *Handbook of Artificial Intelligence for Music: Foundations, Advanced Approaches, and Developments for Creativity* (Springer, 2021), 823–847.

65. T. Grill, *Mutual Understanding*, 2019, accessed 25 February 2026, https://rottingsounds.org/threads/projects/mutual_understanding/.

anical systems, and what approaches to interaction are employed within them.

3.5.1 Current and Emerging Applications

Table 3.1 provides a summary of the key works identified. The seventeen works included are those which use a documented machine learning technique and in which a physical component is central to their realisation. The table categorises each work according to its technique and tool, instrument or physical embodiment, and mode of interaction, providing an overview that informs the creative output of chapters Four, Five and Six.

The ‘Mode of Interaction’ column considers the agents involved, whether an interaction is directional or reciprocal, and the primary artistic purpose (imitative, responsive, generative, etc.). This practical approach was chosen over more formal frameworks, such as those by Jung⁶⁶ or Vear,⁶⁷ to accommodate the diversity of the works, which range from human-machine collaborations to autonomous systems.

MIDI generation was the most widespread application. These works used a range of approaches from creating custom models (Empty Vessels) to fine-tuning existing models (*Fantasie#4*,⁶⁸ *The Light Comes...*⁶⁹). Generative MIDI models were often paired with player piano, likely reflecting the availability of datasets and compatibility of models with MIDI-enabled instruments. Some works in this category used non-musical data such as transcribed speech as input (*Seven Piano Etudes Speaks...*⁷⁰) but the majority used existing piano datasets. Custom-built instruments often repurposed or combined models (*Little Sound Machines*) for the particular affordances of the instrument.

A smaller number of works turned to alternative generative strategies. These included the modelling of environmental systems, as found in Emard’s *Supraorganism*,⁷¹ where a neural network’s predictions of a honeybee colony dictated the behaviour of the installation. In a distinct approach, the emergent behaviour of an analogue neural network was embodied in

66. Matthias Jung, ‘Intelligent Music Performance Systems: Towards a Design Framework’, *Studia Musicologica Norvegica* 49, no. 1 (2023): 28–44, <https://doi.org/10.18261/smn.49.1.3>.

67. Craig Vear et al., ‘Human-AI Musicking: A Framework for Designing AI for Music Co-creativity’, *AIMC 2023*, 2023, accessed 25 February 2026, <https://aimc2023.pubpub.org/pub/zd46ltn3>.

68. Quadrature, *Fantasie 4*.

69. Armitage and Shepardson, ‘Augmenting the expressivity of the Notochord generative midi model for Arca’s “The Light Comes in the Name of the Voice”’.

70. Thompson IV, ‘Creating Musical Scores Inspired by the Intersection of Human Speech and Music Through Model-Based Cross Synthesis’.

71. Emard, *Supraorganism*.

Table 3.1: Overview of surveyed works

Work	ML-technique	ML-model / tool	Data Source	Instrument	Mode of Interaction
Eyes (2024)	Audio classification	Flucoma	Live guitar, playing technique	Motor-augmented upright piano	Performer $\Leftrightarrow$ machine: generative accompaniment
Gioti (2019)	Audio classification	Feed-forward neural network	Performer rhythms, instruments, technique	Servo-motors, multi-percussion	Performer $\Leftrightarrow$ machine: imitation / initiation
Franklin (2024)	Audio classification	MLStar	Performer rhythm	Solenoid-actuated cajon	Performer $\Rightarrow$ machine: imitation
Macionis and Kapur (2019)	Live Parameter Mapping	Wekinator	EEG bio-signal (participant)	Solenoid-actuated singing bowls	Participant $\Leftrightarrow$ machine: biofeedback loop
Emard (2020)	System Modelling	Custom neural network	Honeybee colony	Motors, wires and glass	Autonomous: corpus-based generation (participant responsive)
Spahn and Faubel (2018)	Analogue neural network	Pop-Neuron Oscillator	Network dynamics	Motors, analogue electronics	Emergent behaviour
Quadrature (2021)	MIDI generation	Piano Genie, Polyphony-RNN	Radio signals, classical piano corpus	Player piano	Live data stream $\Rightarrow$ machine: data-driven generation
Nikrang (2024)	MIDI generation	Ricercar	Classical MIDI corpus	Player piano	Autonomous: corpus-based generation
Thompson IV (2022)	MIDI generation	Polyphony-RNN	Speech, MIDI piano corpus	Player piano	Autonomous: corpus-based generation (<i>movements II, IV, VI</i>)
Evans et al. (2025)	MIDI generation	Groove-Transformer	Performer MIDI	Pipe organs	Performer $\Rightarrow$ machine: generative accompaniment (supervised)
Shepardson et al. (2024)	MIDI generation	Notochord	Performer MIDI	Magnetic Resonator Piano	Performer $\Rightarrow$ machine: daily reinterpretation
Gardener and Debicki (2019)	MIDI generation	Custom model	Custom dataset (MIDI)	Mechatronic cello trio	Autonomous: corpus-based generation
Bacon (2018)	MIDI generation	Tensorflow	Rhythm, harmony and melody models	Multiple devices, found objects	Machine $\Leftrightarrow$ machine: generative composition
Bacon (2021)	Speech transcription	DeepSpeech: (speech-to-text)	Human speech (audience)	Bespoke devices, light projections	Audience $\Rightarrow$ machine $\Rightarrow$ machine: data translation and mistranslation
Grill (2019)	Text and language Generation	Custom model	Not documented	Horn loudspeakers	Machine $\Leftrightarrow$ machine: emergent behaviour
Zuiderveld (2024)	Text and language Generation	Mistral AI (LLM)	Pretrained LLM, custom prompts	Speakers / appliances	Participant $\Leftrightarrow$ machine: Conversational interaction
Savery et al. (2021)	Text and language generation	LSTM, WordNet (lyric generation)	Custom lyric dataset (Verse Scraper)	Shimon (robotic marimba)	Autonomous: corpus-based generation

Spahn and Faubel's *Symbolic Grounding*,⁷² based on the spontaneous self-organising properties of such networks.⁷³

Five works used neural networks for data interpretation, with sources ranging from audio to EEG streams. Here, the main affordance of machine learning is its ease of use and robustness relative to other algorithmic approaches. Audio classification was implemented with existing software libraries (FluCoMa,⁷⁴ MLStar⁷⁵), while live parameter mapping often relied on flexible tools such as Wekinator⁷⁶ (*Where is the Quiet*⁷⁷). These models typically formed part of larger algorithmic systems, incorporating performer-machine feedback loops (*Imitation Game*) and stochastic elements to generate electromechanical responses. While the role of the models themselves in these works was largely functional, Bacon's *PROBE II*⁷⁸ used errors in the speech-to-text process as a central theme of the work.

Natural language processing (NLP) is an emerging area, with four works using models ranging from speech-to-text to large language models. Text generation strategies were used to explore machine personality (*Conversations beyond...*⁷⁹), or model non-human forms of communication (*Mutual Understanding*⁸⁰). These focus on language as a narrative or conceptual medium rather than as a structural source for composition, which suggests opportunities for further work in this area.

Other notable gaps include machine vision, which despite the maturity of tools in adjacent fields, remains underexplored in this domain. Similarly, while the potential for embedding inference directly onto microcontrollers is growing with frameworks like TensorFlow Lite⁸¹ and Edge Impulse,⁸² few works currently take advantage of completely embedded machine learning processes.

72. Faubel and Spahn, *Symbolic Grounding*.

73. Shun-ichi Amari, 'Dynamics of pattern formation in lateral-inhibition type neural fields', *Biological cybernetics* 27, no. 2 (1977): 77–87, <https://doi.org/10.1007/BF00337259>.

74. Tremblay et al., *The Fluid Corpus Manipulation Toolbox (v.1)*.

75. Smith and Deal, 'ML.* Machine learning library as a musical partner in the computer-acoustic composition flight'.

76. Fiebrink and Cook, 'The Wekinator: a system for real-time, interactive machine learning in music'.

77. Macionis and Kapur, 'Where is the quiet: Immersive experience design using the brain, mechatronics, and machine learning'.

78. Bacon, *PROBE II, Subaudition*.

79. Zuiderveld, *Conversations beyond the ordinary*.

80. Grill, *Mutual Understanding*.

81. Abadi et al., 'TensorFlow: A system for Large-Scale machine learning'.

82. Shawn Hymel et al., 'Edge impulse: An mlops platform for tiny machine learning', *arXiv preprint arXiv:2212.03332*, 2022, <https://doi.org/10.48550/arXiv.2212.03332>.

3.5.2 Material Constraints

Zareei's *Audiovisual Materialism*⁸³ points to the inseparability of the material, visual and audible aspects of a work. This can be seen clearly in the alien forms of Bacon's *Probe II*,⁸⁴ which are themselves the sound producing bodies. Likewise, the light and sound patterns of Emard's *Supraorganism*⁸⁵ are a direct visualisation of an emergent system, and in player piano works the physical action of keys and hammers are a central part of the aesthetic.

Pigott, in foregrounding the electromechanical mechanism, distinguishes between works which celebrate the unpredictable qualities of materials and those which prioritise fine expressive control.⁸⁶ The most common approach was the latter, with existing instruments including player piano and MIDI-enabled pipe organ, and new augmented instruments such as the Magnetic Resonator Piano⁸⁷ (MRP). The constraints of these systems often influenced the implementation of the machine learning system: in *Empty Vessels*,⁸⁸ Gardener and Debicki created a dataset tailored to the unique affordances of the mechatronic cello trio, finding that existing datasets did not suit the capabilities of the instrument. In Arca, Armitage and Shepardson's *The Light Comes in the Name of the Voice*,⁸⁹ the particular timbral and dynamic affordances of the MRP required a custom approach to model training, reusing aspects of the original performance to map between the expression of the instrument, and the velocity values expected by the model.

Material constraints are however more immediately visible in Pigott's description of unpredictable sound materials. This approach, which Valle terms 'semi-random variations...in the electromechanical environment'⁹⁰ typically features DIY instruments built from motors, servos, solenoids, and found objects. Here, the inherent limitations and erratic nature of the materials act as both affordance and constraint; the unpredictability of Emard's glass-enclosed wires,⁹¹ or ringing school-bell of Bacon's *Little Sound Machines*⁹² are central to those works, but the model and electromechanical mechanism are often less tightly integrated, with the model's output serving as a trigger for an action rather than a detailed performance instruction (as in

83. Zareei, 'Audiovisual materialism'.

84. Bacon, *PROBE II, Subaudition*.

85. Emard, *Supraorganism*.

86. Pigott, 'Across Fields: Sound, art and technology from an electromechanical perspective'.

87. McPherson, 'The magnetic resonator piano: Electronic augmentation of an acoustic grand piano'.

88. Debicki and Gardener, *Empty Vessels*.

89. Armitage and Shepardson, 'Augmenting the expressivity of the Notochord generative midi model for Arca's "The Light Comes in the Name of the Voice"'.
 90. Valle, 'Making acoustic computer music: The Rumentarium project'.

91. Emard, *Supraorganism*.

92. Bacon, *Little Sound Machines*.

the MRP). Where instrumental affordances were particularly constrained—as with solenoid-actuated singing bowls (*Where is the Quiet*⁹³)—a threshold was used to map complex data to physical action. This strategy occurs frequently in wider electromechanical sound art, such as Greg Kress’ *Brain Noise Machine*⁹⁴ which used a threshold on the participant’s level of mental focus to control a single DC motor.

A third distinct approach was the augmentation of existing non-musical appliances, with Zuiderveld embedding personality and narrative into coffee machine, microwave and other devices.⁹⁵ Here, the object’s pre-existing function shapes the artistic statement. This is an idea widely explored in electromechanical sound art outside of machine learning approaches: for example, Dunham’s *CLICK::TWEET*⁹⁶ which draws a comparison between modern and past methods of communication through pairing telegraph keys with social media feeds.

3.5.3 Interaction

The nature of interaction ranged across the surveyed works, from human-machine dialogue to fully autonomous systems. In directional (performer $\Rightarrow$ machine) interactions, a performer’s audio or MIDI data served as a one-way input to the system. The purpose of one-way interactions in the surveyed works ranged from direct imitation to using performance data as a seed for a generative MIDI model. While most works in this category involved live interaction, Arca, Armitage and Shepardson’s *The Light Comes in the Name of the Voice*⁹⁷ presented a delayed interaction in which Arca’s improvisations were reinterpreted by the model in an installation setting.

Reciprocal (performer $\Leftrightarrow$ machine) interactions involved feedback loops. In Parkinson and Dunning’s *Improvising with Machines*,⁹⁸ these systems could be described as collaborative; works in which a human performer responds to the output of the machine, and in turn influences the machine. This dynamic is central to creating what Vear describes as *liveness*, ‘a sensation

93. Macionis and Kapur, ‘Where is the quiet: Immersive experience design using the brain, mechatronics, and machine learning’.

94. Gregg Kress, *Brain Noise Machine*, 2010, accessed 25 February 2026, <https://glkress.com/art-and-design/brain-art/brain-noise-machine/>.

95. Zuiderveld, *Conversations beyond the ordinary*.

96. Paul Dunham, ‘Click:: TWEET. A Telegraphic Sound Installation’, in *Proceedings of the International Conference on New Interfaces for Musical Expression* (2021), <https://doi.org/10.21428/92fbeb44.48c39866>.

97. Armitage and Shepardson, ‘Augmenting the expressivity of the Notochord generative midi model for Arca’s “The Light Comes in the Name of the Voice”’.

98. Parkinson and Dunning, ‘Improvising with Machines: A taxonomy of musical interactions’.

that something is co-operating in the real-time flow, or not’.⁹⁹ Gioti, in *Imitation Game*,¹⁰⁰ describes balancing this real-time interaction with aspects of compositional control by using a naïve rehearsal format, in which a human improviser has no preconception of how the system will respond.

Other works minimised or removed the human from the interactive loop, focusing instead on data-driven or emergent behaviours. Autonomous systems (*Empty Vessels*,¹⁰¹ *Absolute Hallucination*¹⁰²) operated without direct human input during performance. Quadrature’s *Fantasie#4*¹⁰³ presented an interaction between a non-human data source (live radio waves) and a non-human musical agent (generative model, player piano), with the creative act as the machine interpretation of data streams. Emard’s *Supraorganism*¹⁰⁴ combined both dataset-based generation and responsive approaches, through a largely autonomous system, but with audiences influencing the installation’s behaviour through their presence.

Machine-to-machine approaches explored emergent behaviour between multiple machine agents, as in Bacon’s *Little Sound Machines*,¹⁰⁵ where three models ‘learn from and influence each other’. Grill’s *Mutual Understanding*¹⁰⁶ focused specifically on the outcomes of these machine-to-machine interactions in language processing, and Bacon’s *PROBE II*¹⁰⁷ examined interaction through the translation and mistranslation of information, passed between human and machine, and between machines themselves.

Jung proposes a three-axis model for intelligent music performance system design¹⁰⁸ (IMPS) comprised of embodiment, participation, and autonomy—described as the ‘extent to which non-human agents are integrated into a system, and... the autonomous contributions within a system’. The surveyed works as material objects demonstrate high levels of physical embodiment, but vary widely in autonomy and participation. Performer-machine works such as Gioti’s *Imitation Game*¹⁰⁹ featured high participation with autonomy moderated through performer control,

99. Vear et al., ‘Human-AI Musicking: A Framework for Designing AI for Music Co-creativity’.

100. Gioti, ‘Imitation game: Real-time decision-making in an inter-active composition for human and robotic percussionist’.

101. Debicki and Gardener, *Empty Vessels*.

102. Nikrang, *Absolute Hallucination*.

103. Quadrature, *Fantasie 4*.

104. Emard, *Supraorganism*.

105. Bacon, *Little Sound Machines*.

106. Grill, *Mutual Understanding*.

107. Bacon, *PROBE II, Subaudition*.

108. Jung, ‘Intelligent Music Performance Systems: Towards a Design Framework’.

109. Gioti, ‘Imitation game: Real-time decision-making in an inter-active composition for human and robotic percussionist’.

or predefined modes of interaction. Corpus-based systems like Nikrang's *Absolute Hallucination*¹¹⁰ demonstrate a high degree of autonomy by eliminating the human from the immediate interactive loop, even while its output is shaped by a human-curated dataset.

From an audience's perspective, it may not be clear to what extent a work is autonomous.¹¹¹ This is examined in Anna Savery's work for the robot musician Kierzo,¹¹² where an audience survey revealed questions about the whether the device was pre-programmed or responding randomly to the performer—and suggests explaining the technologies before performances as a possible approach. A degree of misinterpretation can also happen in the opposite context: in Stapleton and Davis' networked performance piece *Ambiguous Devices*,¹¹³ pre-concert talks explained that the devices were not autonomous but performed remotely by a human collaborator.

3.6 Conclusion

This chapter finds that the application of neural networks in electromechanical sound art has so far centred on symbolic generation, particularly through MIDI models paired with player pianos and other MIDI-enabled instruments. Beyond this, neural networks have been applied in more functional roles, such as audio or biosignal classification, and in a smaller number of cases to simulate environmental systems or emergent behaviours. NLP, machine vision, and embedded inference remain underexplored in this domain.

Material constraints shaped the surveyed works both in terms of design and function. Those emphasising fine expressive control often adapt datasets or training methods to suit the specific affordances of the instrument, ensuring a close integration between model output and mechanical action. By contrast, DIY instruments often foreground the unpredictable qualities of materials, where limitations and irregularities become aesthetic features. A further strand of practice augments existing appliances, where the pre-existing functions of objects shape the conceptual focus of the work. Within these varied approaches to sound production, a range of interaction approaches were implemented from directional performer-to-machine control, through reciprocal

110. Nikrang, *Absolute Hallucination*.

111. Edmond Couchot, 'Automatism, Autonomy and Aesthetics in the Performing Arts', *Leonardo* 52, no. 3 (2019): 240–246, https://doi.org/10.1162/leon_a_01752.

112. Anna Savery and Richard Savery, 'Collaboration Between Robots, Interfaces and Humans: Practice-Based and Audience Perspectives', *arXiv preprint arXiv:2407.16966*, 2024, <https://doi.org/10.48550/arXiv.2407.16966>.

113. Paul Stapleton and Tom Davis, 'Ambiguous Devices: Improvisation, agency, touch and feedthrough in distributed music performance', *Organised Sound* 26, no. 1 (2021): 52–64, <https://doi.org/10.1017/S1355771821000054>.

feedback systems, to autonomous and machine-to-machine interaction. Directional systems allow performers to seed or guide generative processes, while reciprocal approaches involve feedback loops, where performers shape and are shaped by the machine's behaviour. Autonomous systems minimise human input, drawing on large corpus or non-human data sources, and machine-to-machine interactions explore emergent behaviours between non-human agents.

The survey was not exhaustive, and many other applications and approaches will be in use across the field. The search was limited by the terms and available documentation, with several existing works not including detailed processes. This could be addressed in future through formal interviews with artists.

Overall, the survey presented here offers a partial view of the field, but one that suggests scope for further exploration of how machine learning systems can intersect with electromechanical sound-making. In response to the identification of NLP as an underexplored area, in Chapter 5 I apply speech-trained RAVE audio models within electromechanical practice, with *Smalltalk* developing a method for identifying musical gestures with specified pitch and amplitude contours, and *Radio User Guide* focused on the gradual emergence of semantic meaning from generated audio and text. Chapter 6 instead explores this gap through linguistic patterns, where I apply LSTM text models and a small language model to generate linguistic patterns for musical mapping, in the installation work *Porcelain Music Box*, the improvisatory system for *Seven Studies* and the composition for *Word games*. Approaches to material constraints and improvisational interaction are referenced throughout the following research, beginning with applications of algorithmic speech transcription to new DIY devices.

Chapter 4

Algorithmic Transcription in Practice

4.1 Introduction

In this chapter I develop methods of algorithmic speech transcription through the creation of two DIY motor-driven devices: *Motorcase I* and *Motorcase II*. The work builds on existing practices in speech-derived music by combining algorithmic transcription with live improvisation to convert recorded speech into mechanical sound.

Approaches to instrumental resynthesis discussed in Chapter 2 often involve deliberate approximation, heard in the gradual emergence of speech in Harvey's *Speakings*¹ or the reductive automated transcription of Malherbe's *Locus*.² In this chapter, I extend this approach to approximation by using DIY, low-resolution hardware to explore the abstract rendering of speech. This act draws on Cassandra Miller's concept of *transformative mimicry*, which employs imperfect transcription as a generative process³—applied in this context via approximate computational transcription rather than intuitive listening. The resulting material forms a palette for live improvisation, shaped by a sonic unpredictability that emerges from a tactile relationship with

1. Jonathan Harvey, *Speakings*, 2007, accessed 25 February 2026, <https://www.fabermusic.com/shop/speakings-p410799>.

2. Claudy Malherbe, *Locus*, 1997, accessed 25 February 2026, <https://brahms.ircam.fr/en/works/work/10464/>.

3. Cassandra Miller, 'Transformative Mimicry: Composition as Embodied Practice in Recent Works' (PhD diss., University of Huddersfield, 2018).

found objects and DIY electronics.⁴

This work is guided by R1:

- R1. Which features can be computationally extracted from the speech signal, and how can they be implemented in an electromechanical improvisation and composition practice?

In contrast to approaches prioritising intelligibility, such as Canuto’s *Talking Motors*,⁵ or those exploring the drift between intelligibility and abstraction as found in Ablinger’s *Quadraturen* series,⁶ in both *Motorcase I* and *II* I develop transcription methods focused on capturing the general contour and rhythm of speech as a hidden source of musical material. While the projects differ in several aspects, the fundamental difference is in the scale of this source (Table 4.1). In *Motorcase I* I begin with six speech recordings, a small set that enabled a manually-assisted transcription method. Limitations in this approach led to *Motorcase II*, designed to navigate a large-scale corpus of over 100,000 emotionally-labelled speech files. Throughout this chapter, these differences in scale shape the transcription methods, the design of the devices, and my approach to improvisation and composition.

Table 4.1: Overview of Project Elements: *Motorcase I* and *Motorcase II*

	Motorcase I	Motorcase II
Speech Source	6 speech recordings	100,000 speech recordings
Speech Features	Text, spectral centroid, amplitude	Mel-spectrogram, emotion labels
Device	8 DC motors, mixed found objects	32 servo motors, metal found objects
Creative Output	· Improvisation · Fixed composition (video, ensemble)	· Improvisation · Fixed composition (tape, ensemble)

Objectives

Three key objectives of the projects discussed in this chapter are as follows:

1. **Devices:** Create two DIY devices for computer-controlled acoustic sound production, using motors and found objects.

4. John Richards, ‘Getting the hands dirty’, *Leonardo Music Journal* 18 (2008): 25–31, <https://doi.org/10.1162/lmj.2008.18.25>; Andrew Brown and John Ferguson, ‘DIY musical instruments: From Handmade Electronic Circuits to Microcontrollers and Digital Fabrication’, *Journal of Ubiquitous Music*, 2024, accessed 25 February 2026, <https://hdl.handle.net/10072/434542>.

5. Jesus Canuto, *Talking Motors*, 2023, accessed 25 February 2026, <https://www.jesuscanuto.com/talking-motors>.

6. Peter Ablinger, *Quadraturen*, 1995 -, accessed 25 February 2026, <https://ablinger.mur.at/docu11.html>.

2. **Speech Transcription:** Develop two strategies in response to the features of each source, avoiding direct reproduction of speech in favour of an abstract rendering.
3. **Improvisation and Composition:** Design interfaces for each device, balancing automated playback with real-time control. Extend generated materials into fixed works for instruments and electronics.

4.2 MotorCase I

4.2.1 Overview

The first of two works in this chapter, *MotorCase I* investigates algorithmic speech transcription through a small selection of speech audio recordings. Existing approaches to instrumental resynthesis identified in Chapter 2 often take a spectrogram as a starting point.⁷ Here, I wanted to begin with a different set of features—by using a text transcription as an alternative method of rhythmic segmentation.

Approaches to the electronic manipulation of speech also described in Chapter 2, such as looping⁸ or altering speed,⁹ prompted the idea that these could be applied to the electromechanical rendering of speech audio, through some kind of live performance interface. For me, and like many experimental and DIY approaches,¹⁰ exploring this idea through an iterative and improvisational approach was the only way to build, test and perform with the system.

To support this approach, and to see how the varied character of differing voices might survive the transcription and mechanical realisation process, I chose a small set of speech audio recordings as the foundation of the system. These recordings form the basis for a portable motor-controlled percussion device, constructed from found objects and embedded within a hard-shell case.

7. Clarence Barlow, *Orchideæ Ordinariæ*, 1989, accessed 25 February 2026, <https://www.soundohm.com/product/musica-algorithmica-2-cd>; Thomas Hummel, *Nicanor*, 1996, accessed 25 February 2026, <https://www.thomashummel.net/en/music/nicanor/>; Andrew Chen, 'Voice Without Speaker: human speech synthesis in acoustic instrumental contexts', *Tempo* 77, no. 305 (2023): 60–71, <https://doi.org/10.1017/s0040298223000098>; Beat Furrer, *Wüstenbuch*, 2009, accessed 25 February 2026, <https://www.kairos-music.com/cds/0013312kai>.

8. Blablabor, *hirsch hirn hornisse*, 2009, accessed 25 February 2026, <https://www.blablabor.ch/hirsch-hirn-hornisse>; Bernhard Gál, *bestimmung new york*, 1999, accessed 25 February 2026, <https://bernhardgal.com/bestimmungnewyork.html#info>.

9. Clarence Barlow, *fLvXvS*, 1990, accessed 25 February 2026, <https://clarlow.org/compositions-by-year>.

10. Andrea Valle, 'Making acoustic computer music: The Rumentarium project', *Organised Sound* 18, no. 3 (2013): 242–254, <https://doi.org/10.1017/s1355771813000216>; Richards, 'Getting the hands dirty'.

Objectives

1. **Source:** Choose a small set of speech recordings to highlight the expressive character of each voice.
2. **Device:** Create a noise-based palette using found objects and actuators, and house the assembly in a portable case.
3. **Transcription:** Use three features as the basis of the speech transcription—text segmentation, the spectral centroid, and amplitude.
4. **Improvisation:** Design an interface for real-time playback and manipulation of each transcription.
5. **Composition:** Apply the methods and device to a new dataset, in a composition for video, tape, and instruments.

4.2.2 Speech Source

Speech sources for *Motorcase I* were limited to six recordings, each approximately one minute in duration. This allowed the project to focus on the characteristics of individual voices, in contrast to the large-scale dataset of *Motorcase II*.

At the outset, the precise results of the transcription process were unknown, but it was clear that the combination of algorithmic mapping and DIY found objects would strip the result of linguistic meaning. Therefore, my primary selection criteria was sonic variation; I chose audio from a spectrum of vocal delivery styles from casual conversation to lectures, through an intuitive assessment of rhythmic character, pitch contour, and timbre. A varied palette at this stage would allow a broader range of material to be developed later in an improvisational context.

The six files, sourced from online archives and personal interviews, included:

1. *Conversational 1*. Low, slow, softly spoken.
2. *Conversational 2*. High, softly spoken, hesitant.
3. *Lecture 1*. Confident, deliberate.
4. *Lecture 2*. Fast and rhythmically irregular.
5. *Ted Talk*. Wide melodic and dynamic variation.

6. *Televangelist*. Fluent, monotone, rhythmically regular.

These recordings provided the foundation for the system. The following sections detail three layers built on this foundation: the physical layer of the device, the algorithmic layer of the transcription process, and the interface for live improvisation.

4.2.3 Device

The function of *Motorcase I* was to realise recorded speech transcriptions through motors and found objects. In the design of the physical device I extend the approximate approach to transcription; as algorithmic transcription transforms speech into an imperfect representation, the instrument adds its own layer of physical instability. This principle of instability applies to the device's construction, which is transparently rudimentary, and held together with glue and tape. The motor-object assembly is built into a portable case, to support live performance (Figure 4.1).

I chose the hardware setup (Figure 4.2) for simplicity, and compatibility with the software used in the project. Acoustic sound was captured with two contact microphones fixed to the case, to amplify details produced by the physical interaction of the motors and objects.



Figure 4.1: *Motorcase I* device, showing the found objects and actuators

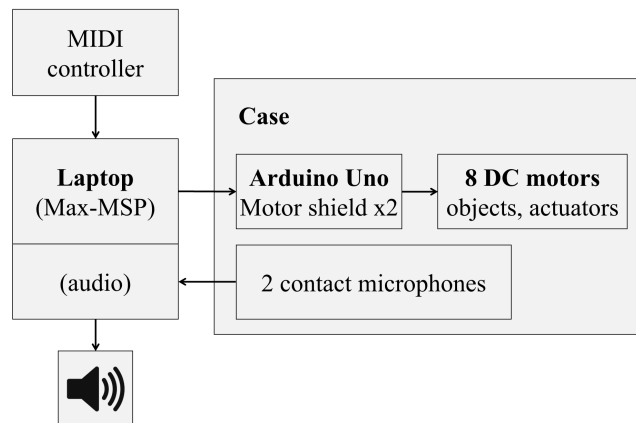


Figure 4.2: *Motorcase I* hardware

For the device itself I used eight DC motors. This provided a range of possible object-actuator configurations while remaining manageable from a control and a portability perspective. The specific objects were then chosen to create a palette of distinct, noise-based sounds. Each pairing

was chosen to have its own character and mode of action: some produced sharp, percussive sounds (bells, ratchet), others created sustained textures (pebble on stones, spoon on grill), and several introduced a degree of chance into the performance.

Portfolio Example 4.1 — *Motorcase I*: Objects and Actuators^a

1. Large pebble rotating across a set of flat stones. ►
2. Terracotta plant pot pressed against a spring-mounted ball bearing. ►
3. Small bells struck laterally by a rubber-band-suspended ball bearing. ►
4. Metal wire scraped laterally by a threaded brass rod. ►
5. Metal grill scraped by a rotating metal spoon. ►
6. Small linked chain moving against a metal wire. ►
7. String-suspended beads rotating against glass jars ►
8. Ratchet mechanism clicking against a hollow metal tube. ►

^a. Portfolio/Chapter 4 - Motorcase I/Examples/...

4.2.4 Transcription Method

My aim for the transcription process was to capture something of the rhythm and contour of speech, rather than its linguistic content. The use of a small dataset afforded a manually-assisted approach, combining spectral analysis with a novel text-based segmentation to produce a palette of material for live improvisation.

Guided by the principle of imperfect transcription as a generative process, approximations of sonic qualities, including rhythm, spectral centroid, and amplitude, serve as control data for the motors. The rhythmic segmentation was the basis for this process, and involved automatically segmenting the audio according to a manual text transcription—letters, words, and longer periods of sound and silence—to create rhythmic patterns that were related but distinct from the original speech (Figure 4.3). This approach was also a response to the small dataset, generating three rhythmic layers from each source file.

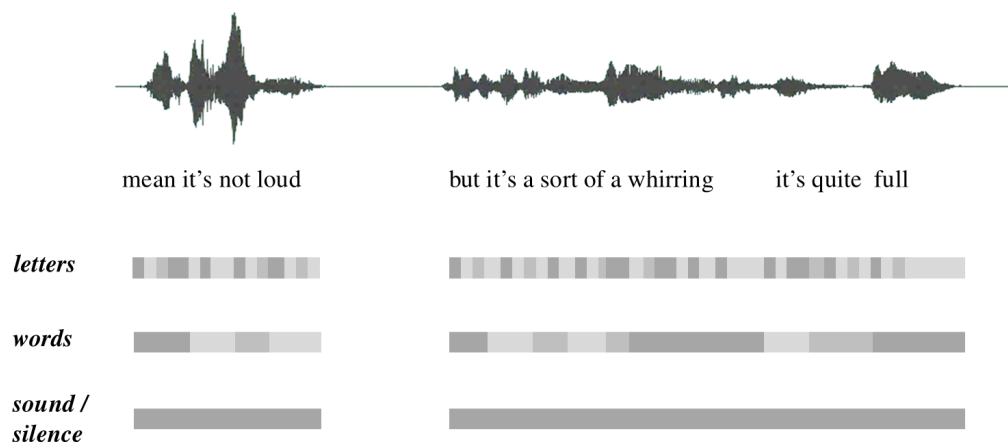


Figure 4.3: Visualisation of the text-based rhythmic segmentation

This idea of variable rhythmic segmentation is related to Ablinger's grid-based *Quadraturen* method,¹¹ in that musical variation is generated through the resolution of a transcription. Unlike Ablinger's externally imposed grid, this approach derives the segmentation from a text transcription of the audio, meaning the resolution of the transcription is internally derived from the source material.

The transcription was precomputed, creating a fixed body of material that could be played back and manipulated in live improvisation. The logic was a direct translation of features to parameters, converting each rhythmic layer, associated spectral centroid, and amplitude into instructions for the performance interface. The resulting material was an abstract rendering of each speech source, providing a palette of sounds for an improvisational practice based around navigating the different layers of the transcription.

Code Example 4.1 — *Motorcase I*: Transcription Detail

A Python script^a transcribed speech audio to motors. Although automated speech-to-text was tested, it tended to remove hesitations, half-words, and repetitions which were essential for capturing rhythmic detail.

1. A text-based rhythmic segmentation was derived from manual text transcription of

11. Ablinger, *Quadraturen*.

the speech audio. Words and letters were automatically aligned with the audio using WAV2VEC2,^b with onsets and offsets returned for these layers. Regions of sound and silence were determined using a simple amplitude threshold.

2. The spectral centroid of each speech segment determined which motor combination was activated, which I chose as a reductive approximation of vocal contour and timbre. The script matched the segment's centroid to the closest pre-calculated centroid from a set of all possible motor combinations.
3. The amplitude of each speech segment was mapped directly to the speed of the selected motors.

a. Code/Chapter 4 - Motorcase I code/TranscriptionMethod.py

b. Alexei Baevski et al., 'wav2vec 2.0: A framework for self-supervised learning of speech representations', *Advances in neural information processing systems* 33 (2020): 12449–12460, <https://doi.org/10.48550/arXiv.2006.11477>.

4.2.5 Improvisation

To play the transcribed speech in performance, I created a Max MSP (Cycling '74) interface as the bridge between the static transcriptions and real-time improvisation. The design was a response to two factors: the segmentation of the speech, and the small scale of the dataset. The interface therefore was structured around two modes of interaction: navigating the transcribed speech data, and manipulating the output.

Navigation controls engaged with the results of the transcription process, moving between the different rhythmic layers (letter, word, and sound/silence). Manipulation controls applied conventional electroacoustic techniques. These features, of looping, speed control, quantisation, and stochastically mixing the six transcriptions were conceptually independent of the speech source. They served to extend live improvisation beyond the mapped data, as raw material to be shaped musically.

Code Example 4.2 — *Motorcase I*: Interface Detail

Individual and global controls within the Max interface^a are shown below (Figures 4.4 and 4.5).



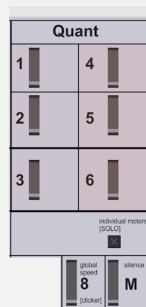
- *1 male convo* refers to the speech sample.
- *Loop* displays start and end time in the sample, adjusted via the two sliders.
- *Time* displays current time in sample.
- *Output* displays current active motors

- *Sound-word-letter* slider adjusts mix between different rhythmic layers.

- *On* unmutes the sample.
- *reset* clears all adjustments.

- *Mix* slider adjusts the mix of the sample in the overall mix.
- *[stones]* refers to the object in direct control mode.

Figure 4.4: Individual sample controls



- *Quant* sliders adjust the quantisation of each sample individually.
- *individual motors [SOLO]* changes mode to direct control
- *global speed* adjusts the overall playback speed for all samples.
- *silence M* adjusts the likelihood of a given sample being replaced by silence.

Figure 4.5: Global controls

^a Code/Chapter 4 - Motorcase I/Code/ImprovPatch.maxpat

I found that improvising with *Motorcase I* involved a balance between presenting the transcribed data, shaping it via the interface, and responding to the algorithmic and physical unpredictability of the device. Through practice, I learned the character of each speech mapping; its rhythmic qualities, density, and contour. Familiarity allowed an intentional approach to performance where I could guide the system to specific areas, for example, by selecting a sparse section to loop or mixing two rhythmically contrasting samples.

This intentionality was at odds with the system's stochastic behaviour and the physical instability of the objects. As a result, a key part of the improvisational process was waiting for a moment to capture, loop, and develop, with each performance exploring a different aspect. While the relationship between movement and sound was transparent, the underlying speech-derived process was hidden. The work itself presents the results, rather than a transparent presentation of the process. In the composition, I take a different approach, in which video and speech audio make visible the connection between source, process, and output.

Portfolio Example 4.2 — *Motorcase I: Improvisation in Practice*

The *Portfolio* shows an example of duo improvisation with Gaia Blandina.^a ► Extracts relating to the interface are highlighted below.

- Density adjustment at 0:37, to create a sparse texture.
- Allowing the device to play at 01:12 without any adjustments, waiting for material to capture and modify.

- Object-unpredictability at 01:38 as the spoon reaches the hanging bell.
- Looping adjustments at 05:58 to create variation within a simple loop. The exact response of the device is unknown, but the basic loop remains the same.
- Manual control at 04:38 and 08:00 to create space within an otherwise dense texture, and highlight the sonic qualities of individual objects. At 10:56 manual control is used to create a dense combination of all objects.

a. Portfolio/Chapter 4 - Motorcase I/Motorcase I improvisation

4.2.6 Composition

To explore how the system could function with a new dataset, and in dialogue with instruments and electronics, I applied the device and transcription methods to a fixed composition. *Years a Have to Know or Something* (2023) was written for *Motorcase I*, bass clarinet, guitar, violin, cello, and video/tape. The work explores the themes of mechanisation, linguistics, and communication technologies, collaging video clips from archive documentaries and films.

Unlike live improvisation with the device, the tape part created a link between recorded speech and the device—a strategy used in several works identified in previous chapters to reinforce links between speech and mapped results.¹² Like these examples, I alternate clips of video and transcribed motors to make this connection, layering instrumental parts devised through automated and intuitive transcription.

Code Example 4.3 — *Motorcase I*: Video Detail

Video extracts were triggered by MIDI input in a Max patch^a with motor instructions stored in `p[Motor Output]`. Video clips were selected from archives, chosen for their visual, linguistic, or humorous qualities:

- Phonetic sounds taken from documentaries and educational films, including Lewis Thomas' 1965 *Normal Speech Articulation*.

12. Henrik Hellstenius, *Instrument of Speech*, 2016, accessed 25 February 2026, https://hellstenius.no/wp-content/uploads/2018/09/PLING_An-expanded-palette-2.pdf; Phillip Venables, *Answer Machine Tape, 1987*, 2022, accessed 25 February 2026, <https://philipvenables.com/2022/09/25/answer-machine-tape-1987>; Matthew Grouse, *Left Right, Left Right*, 2020, accessed 25 February 2026, <https://www.matthewgrousemusic.com/left-right-left-right>; Ondrej Adámek, *Ça tourne ça bloque*, 2007, accessed 25 February 2026, <https://ondrejadamek.com/listen/ca-tourne-ca-bloque-10-musicianssamplevideo/>.

- Clips themed around speech and communication used for visual or humorous purposes, taken from educational films including McGraw-Hill's 1957 *Is There Communication When You Speak?*.
- Technical demonstrations including of the telephone, phonograph and other recording technologies, from documentaries such as Bell Telephone Lab's 1968 *The Incredible Machine*, and Fleischer's 1947 *Mr Bell*.
- Visions of a technological future including the 1936 National Association of Manufacturers's *Men and Machines*.

a. Code/Chapter 4 - Motorcase I code/Video composition patch.maxpat

Each video clip was run through the transcription process, and edited by hand, an affordance of writing a fixed composition. Here, the limitations of the transcription method were a central feature. This informed my approach to the instrumental writing, where I used a variety of unstable and noise-based instrumental techniques linked to the qualities of the *Motorcase I* device. These included dampened strings, bouncing bow, extreme sul pont and sul tasto, harmonics, scraping the guitar fingerboard, small glissandos, clarinet key clicks and mixtures of air and pitch.

Portfolio Example 4.3 — *Motorcase I*: Video Composition

Extracts relating to the compositional approach in the video^a ► and score^b ≡ documentation are highlighted below.

- Unstable instrumental techniques to mimic the qualities of the found objects and motors, found in bar 2, 13, 96.
- Instrumental approximations reflect the approximate nature of the motor transcriptions, outlining the general contour of the speech at 01:58 ("and the speech itself") or 02:03 ("my").
- Clips chosen for rhythmic qualities at 03:58, where syllable stress patterns are the focus.
- Clips used for humorous, theatrical or visual effect at 03:57, 04:21, 05:05 and elsewhere to delineate sections, or 05:56 for the striking image.

- Transparent process beginning at 06:00 presents the gradual replacement of speech with motors, in a playful reference to the meaning of the clip as the machine fails to reproduce speech in any intelligible way.

a. Portfolio/Chapter 4 - Motorcase I/Years a have to know or something.mov

b. Portfolio/Chapter 4 - Motorcase I/Years a have to know or something score.pdf

4.2.7 Evaluation

In the following evaluation I consider the sequence of translation from the source recordings to the physical output, exploring how each stage contributed to the overall goal of an abstract rendering of speech through motors and found objects. This reveals a tension between preserving the expressive character of the source, and achieving a form of transformative mimicry through the transcription and realisation.

Source to Transcription

I found the selection of six speech recordings provided a small but diverse dataset for the transcription process. This is evident in Figure 4.6, where variation in speaking rate and rhythmic regularity across recordings are shown through the first 10 seconds of speech onsets. This also confirmed an intuition that the regularity of the *Televangelist* speech—relative to the other sources—was captured by the transcription process.

The text-based segmentation generated three layers of musical material from each transcription, demonstrating the effect of a text-based resolution (Figure 4.7). In particular, multi-syllable words were condensed into single onsets in the word-level segmentation, which is an expected consequence of this representation. Likewise, individual letters generated multiple onsets for short words—for example, ‘was’ becomes a three-onset figure (w-a-s). These effects were predictable outcomes of the segmentation strategy, but illustrate how each level of abstraction shapes the resulting rhythmic material.

As this created new musical material, it moved the output further from the prosody of each source. A significant amount of individual vocal character was lost during the mapping process, an inevitable consequence of translating continuous spectral and amplitude features into only eight motors and discrete amplitude classes—a coarse quantisation, which flattened much of the nuance of the original voices.

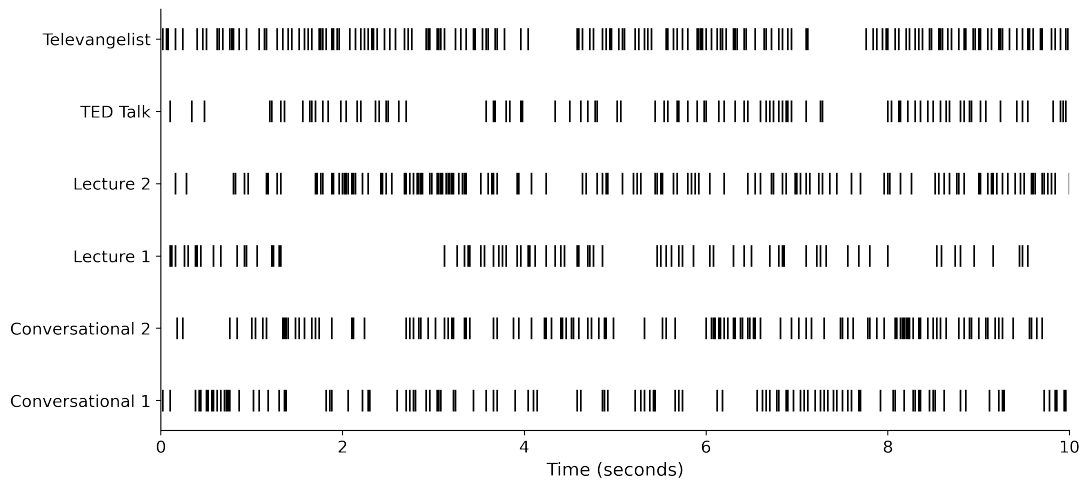


Figure 4.6: Plot of speech onsets within the first 10 seconds of each transcription, showing variation in density and spacing.

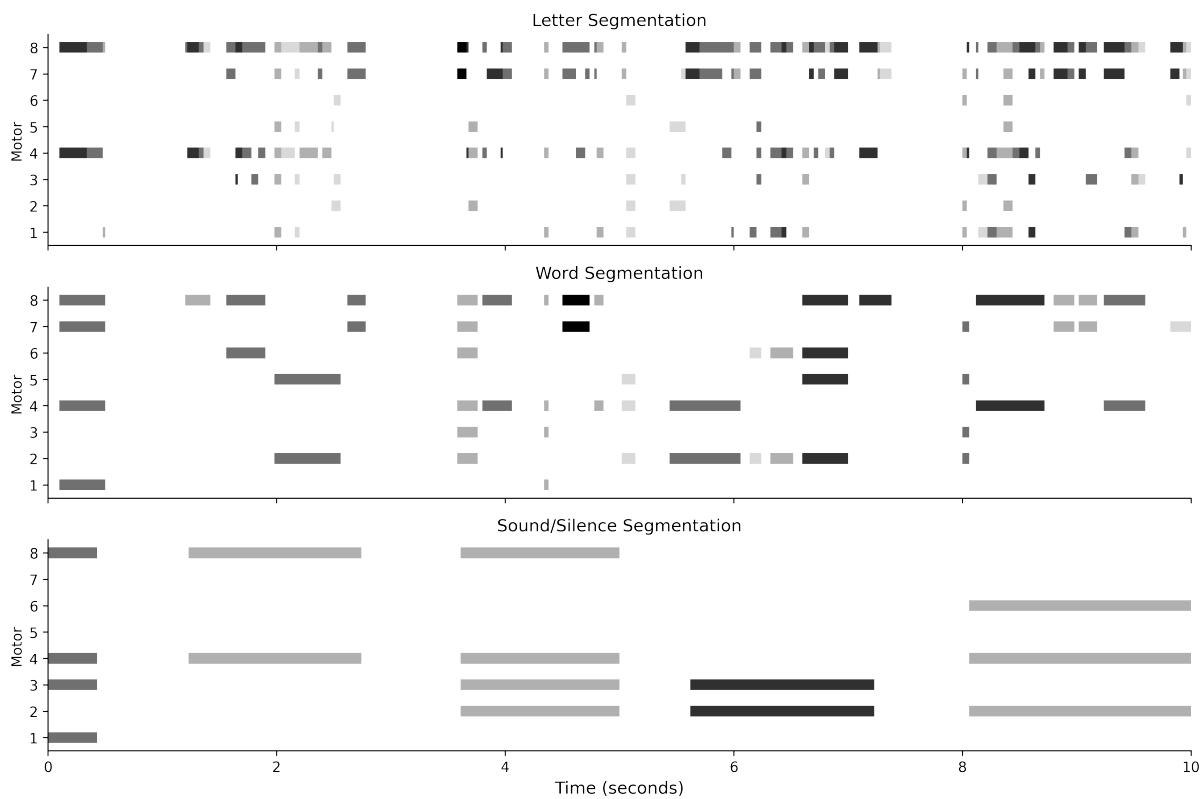


Figure 4.7: Plot of transcription data within the first 10 seconds of the 'Ted talk' audio, showing the levels of the text-based segmentation.

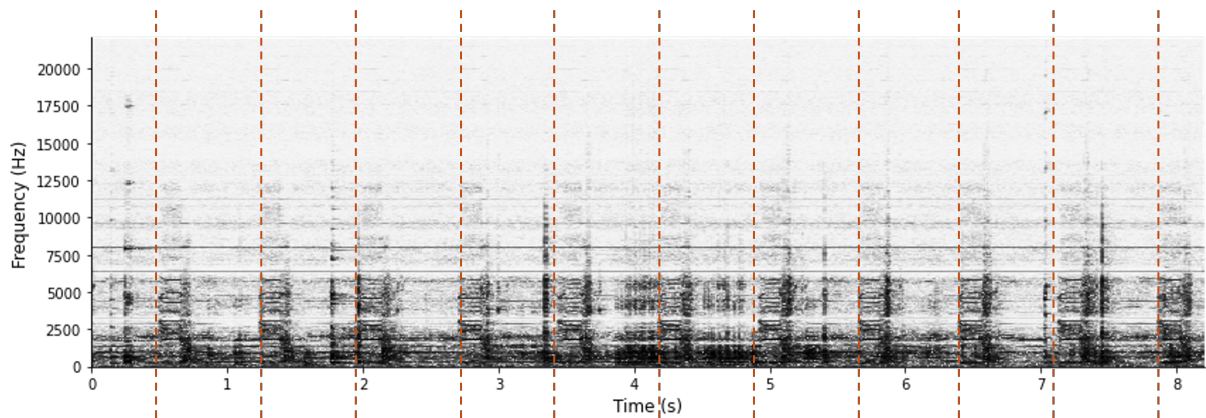


Figure 4.8: Spectrogram of short looping sequence on the device, showing variation due to variability in the object-actuator configurations. Loop marked with dashed orange lines.

Portfolio Example 4.4 — *Motorcase I*: Transcription Examples^a

Examples of the motor-mapping in comparison with speech audio show the loss of information in the ‘words’, and ‘sound and silence’ segmentations, while the ‘letters’ segmentation adds rhythmic information not present in the original speech audio.

- **9 Letters** *Lecture 1* ►
- **10 Words** *Lecture 1* ►
- **11 Sound and Silence** *Ted talk* ►

^a. Portfolio/Chapter 4 - Motorcase I/Examples/...

Transcription to Device

The translation from transcription data to mechanical realisation introduced a second, physical layer of approximation. The first aspect was the instability of the object-actuator configurations, such as the spoon’s interaction with the bell detailed previously, or variation in simple loops caused by the variability of each object (Figure 4.8). For me, this aspect introduced a lightness and humour to each performance, a sense—of course metaphorical—that the device had its own ideas about where the music should go.

Compounding the instability of these object-actuator configurations, the hardware itself created discrepancies—for example, the DC motors were often unable to execute rapid, short-duration commands, causing omitted data or a ‘blurring’ effect between adjacent segments. This becomes clearer when transcribing by ear the speech rhythm, the motor data, and the

perceived rhythm produced by the device (Figure 4.9). These differences were part of the project’s aesthetic; I see the transformative mimicry as occurring both in the gap between audio and transcribed data, and between the data and the sound produced by the device. Together, they result in an output transformed from the source material, but retaining traces of the original speech.

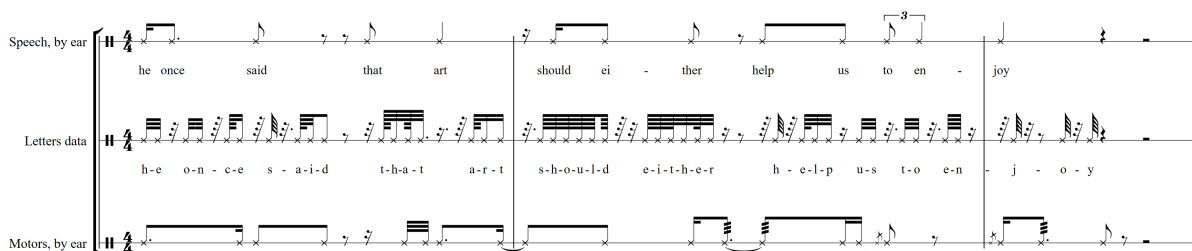


Figure 4.9: *Motorcase I* transcription showing differences between speech, motor data, and motor realisation in the letters segmentation

Improvisation and Interface

The interface was designed to bridge the static transcriptions and live performance. In practice, the interface controls—such as looping, density adjustment, and speed control—functioned as tools for exploring and capturing fragments of the transcribed data. I found that this level of intervention introduced a tension between data and performance: manipulation through the interface often resulted in an aesthetic decoupling from the source. This appeared to mask the traces of prosodic character which survived the mapping process, and shift the focus to the stand-alone material qualities of the device. In the case of live improvisation, I did not see the visibility of the original speech source as essential for performing with the device.

Composition

In applying the system to a fixed composition, I did intend the link between mechanical sound and speech source to be visible. However, this highlighted some limits of the transcription method. In *Years a Have to Know or Something*, the automated mapping required manual editing to fully integrate the generated materials into the fixed composition; while the approximate transcription was effective for generating raw material for improvisation, I felt the transcription stage sat awkwardly in a compositional process where multiple rounds of editing and refining

were required. This motivated the next stage of research, towards an automated transcription method in *Motorcase II*.

4.2.8 Summary

In *Motorcase I* I developed a practice research project involving the abstract rendering of speech into mechanical sound. A small set of speech recordings were selected, and transcribed to motors using text segmentation, the spectral centroid and amplitude. Discrepancies between the source and realisation contributed to the effect of transformative mimicry, an approximate transcription generating novel musical material.

An interface was developed for live improvisation with the transcriptions, balancing automation with real-time control. Here, the strategies I developed for *Motorcase I* were shaped by the scale of the dataset. The interface, with its focus on looping and mixing, was a necessary tool to generate variety from a small pool of data. While this constraint helped in developing a familiarity with the source material, the musical possibilities were limited by the six pre-analysed files, and the scale of the manipulations often masked the remaining traces of speech prosody.

The transcription method and device were applied to a fixed composition for video and small ensemble, in a work which clarified the link between source speech and motor output. In composition, the need to manually edit the automated transcription for each clip suggests that while the method was a useful starting point, it lacked the convenience required for a fixed work. Any such manual process would be unscalable for a system dealing with a larger dataset.

This motivated the next stage of the research, which moves from manipulating a few known samples to navigating a vast corpus, based on automated transcription. This required a method that could work on a larger scale, and an interface designed for discovering material within this corpus, set out in the following project.

4.3 Motorcase II

4.3.1 Preamble

This research often involved experimentation and failure, with unfinished works informing larger projects. One precursor to *Motorcase II* was an abandoned experiment in emotional speech transcription, which consisted of objects fixed to a wooden board.

Instead of direct speech transcription, I used a Markov model to emulate the statistical rhythmic and dynamic qualities of emotional speech. However, the project was abandoned for two reasons: the musical output proved limited, and the physical device was unreliable.

The failure of this approach reinforced the value of direct transcription, and the portability of *Motorcase I*. The following work, *Motorcase II*, builds on the speech transcription-based methods of the first device with a large-scale emotional corpus, using the same motors as the precursor in a more robust and performable device.

Portfolio Example 4.5 — *Motorboard: Video*^a

While the abandoned device ► tended towards self-destruction, this was unintentional. Deliberate explorations of such behaviour exist elsewhere within kinetic art.^b

^a. Portfolio/Chapter 4 - Motorcase II/Preamble/MotorBoard.mp4

^b. Christina Chau, 'Movement and Time in the Nexus between Technological Modes with Jean Tinguely's Kineticism', in *Arts*, vol. 3, 4 (MDPI, 2014), 394–406, <https://doi.org/10.3390/arts3040394>.

4.3.2 Motorcase II: Overview

In the second and final work of this chapter, I address R1 through the implementation of speech transcription at a large scale. Chapter 2 identified speech corpora as an underexplored source of musical material; in response, I move beyond the small dataset of *Motorcase I* to the vast MSP-Podcast¹³ dataset of 100,000 audio files. The resulting device, *Motorcase II*, is a self-contained system for live improvisation, expanded to thirty-two motors and an array of found metal objects.

The dataset contains emotional metadata in the form of valence, activation and dominance labels¹⁴ (VAD), which describe the emotional quality of the voice. I control the device itself through these labels, providing a novel way to navigate a speech dataset in real time: control is approximate, indicating an emotional region, and the system responding by retrieving a pre-transcribed file from that region, and playing it through the motors. In this process, from the approximate guidance to the material realisation, I draw on the recurring theme of transformative mimicry,¹⁵ in which imperfect transcription becomes a generative act.

Objectives

1. **Source:** Choose a large dataset of recordings, to focus on the navigation of a speech corpus.
2. **Device:** Create a homogenous palette of found metal objects. Improve portability by using a Raspberry Pi and integrated amplification.
3. **Transcription:** Use the Mel-spectrogram for the transcription of speech to motors, to capture something of the rhythmic, timbral, and dynamic qualities of emotional speech in a single process.
4. **Improvisation:** Develop an interface for navigating the corpus in real-time, using emotional labels as primary control.
5. **Composition:** Apply the device to fixed composition for small ensemble and tape, using

13. Reza Lotfian and Carlos Busso, 'Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings', *IEEE Transactions on Affective Computing* 10, no. 4 (2017): 471–483, <https://doi.org/10.1109/TAFFC.2017.2736999>.

14. James A Russell and Albert Mehrabian, 'Evidence for a three-factor theory of emotions', *Journal of research in Personality* 11, no. 3 (1977): 273–294, [https://doi.org/10.1016/0092-6566\(77\)90037-x](https://doi.org/10.1016/0092-6566(77)90037-x).

15. Miller, 'Transformative Mimicry: Composition as Embodied Practice in Recent Works'.

RAVE¹⁶ audio models to clarify the link between speech source and mechanical sound.

4.3.3 Speech Source

Recent developments in machine learning have prompted the creation of several large speech datasets, including in the field of emotion recognition.¹⁷ While the musical mapping of emotional speech remains relatively unexplored, the exaggerated sound qualities,¹⁸ as well as research suggesting shared acoustic cues between emotional speech and music—such as speech rate or tempo, vocal intensity or sound-level variability¹⁹—point to emotional speech as a uniquely expressive source for musical mapping.

Many datasets use acted speech consisting of repeated phrases,²⁰ but in this project I chose a corpus of naturalistic emotional speech to broaden variation in the output of the device. A key objective was to navigate the dataset using emotional labels, and so I excluded those which used core emotion labels (happy, angry, sad etc.) in favour of a dataset labelled with valence (pleasant to unpleasant), activation (low-high energy), and dominance (degree of control). This led to the selection of the MSP-podcast dataset, a vast corpus of over 100,000 files of naturalistic speech, using a three-axis VAD scale with each axis ranging from one to seven, in which the emotion of ‘angry’ would broadly correspond to low valence, high energy, and high dominance. These VAD labels provided a three-dimensional space through which to explore the dataset.

Code Example 4.4 — *Motorcase II: Dataset and Navigation*

A Python script^a retrieved the nearest audio file to a given VAD coordinate, through the function `find_closest_filename_optimized()`.

^a. Code/Chapter 4 - Motorcase II code/Improvisation.py

16. Antoine Caillon and Philippe Esling, ‘RAVE: A variational autoencoder for fast and high-quality neural audio synthesis’, *CoRR*, 2021, <https://doi.org/10.48550/arXiv.2111.05011>.

17. Monorama Swain, Aurobinda Routray and Prithviraj Kabisatpathy, ‘Databases, features and classifiers for speech emotion recognition: a review’, *International Journal of Speech Technology* 21 (2018): 93–120, <https://doi.org/10.1007/s10772-018-9491-z>.

18. Dimitrios Ververidis and Constantine Kotropoulos, ‘Emotional speech recognition: Resources, features, and methods’, *Speech communication* 48, no. 9 (2006): 1162–1181, <https://doi.org/10.1016/j.specom.2006.04.003>.

19. Patrik N Juslin and Petri Laukka, ‘Communication of emotions in vocal expression and music performance: Different channels, same code?’, *Psychological bulletin* 129, no. 5 (2003): 770; Eduardo Coutinho and Nicola Dibben, ‘Psychoacoustic cues to emotion in speech prosody and music’, *Cognition & emotion* 27, no. 4 (2013): 658–684, <https://doi.org/10.1080/02699931.2012.732559>.

20. Steven R Livingstone and Frank A Russo, ‘The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English’, *PloS one* 13, no. 5 (2018): e0196391, <https://doi.org/10.1371/journal.pone.0196391>.

4.3.4 Device

In building the *Motorcase II* device I intended to mirror something of the change in the scale of the source speech, with an expanded set of thirty-two motors allowing for a detailed mapping of the speech audio. While the construction was more robust than its predecessor, the guiding principle of sonic instability remained, and the interaction between the actuators and objects ensured the system's realisation of the data was always variable.

One objective was to improve on the portability of the first device by creating a self-contained system, removing the need for an external laptop. To this end, control and amplification were integrated into a portable case (Figure 4.10 and 4.11).

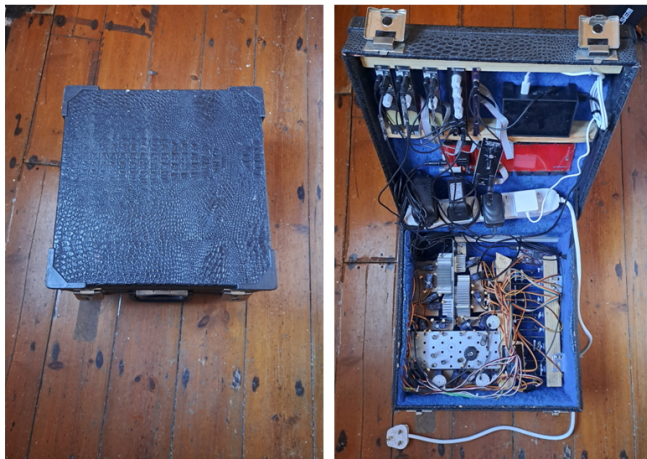


Figure 4.10: *MotorCase II*, open and closed.

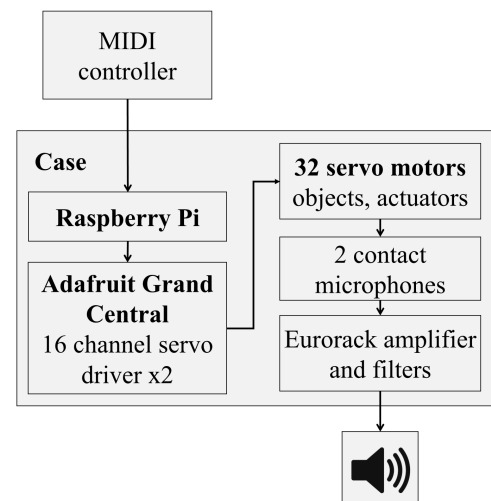


Figure 4.11: *MotorCase II*, hardware

Unlike *Motorcase I*, I made an aesthetic choice to create a more consistent timbral palette, by limiting the sound sources to found metal objects. These were selected through an intuitive process, from objects to hand, and included computer heatsinks, springs, toy piano bars, and small metal sheets. Actuators were simple constructions of rubber bands, springs, wooden beads, and metal washers, arranged to introduce controlled unpredictability into the realisation of the speech transcriptions. For example, the changing positions of rotating beads would strike objects differently when triggered, and the exact point of contact between a spring and toy piano bar would vary slightly with each movement.

4.3.5 Transcription Method

The scale of the dataset required moving from the manually-assisted process of *Motorcase I* to an automated method. Still, my goal remained to balance the representation of speech with its transformation through the transcription process and the device's physical materials.

Since the device was not pitched in a traditional sense, an absolute pitch mapping seemed unsuitable. Instead, to capture something of the relative speech contour, I chose a thirty-two band Mel-spectrogram analysis (corresponding with the thirty-two motors of the device). This created a direct relationship between a frequency region and motor, and allowed the mapping to be reconfigured during performance as a strategy for creating musical variation. The analysis of each audio file was converted into a corresponding MIDI file, creating a large corpus of playable material indexed by the emotional labels of the dataset.

The advantage of using the Mel-spectrogram as the speech feature was in capturing both pitched and non-pitched elements of speech, unlike fundamental frequency tracking which miss noisy or whispered vocalisations.²¹ Also, unlike pre-trained machine learning models for pitch transcription (such as Spotify's *Basic Pitch*²²), this method allowed finer control and iterative adjustment of the mapping process.

Code Example 4.5 — *Motorcase II*: Transcription Preprocessing

For each time-step of the audio, the script:^a `</>`

- Identified the top n mel-bands with the highest energy, and ignored those below a minimum threshold. This allowed control over the density of the mapping, while ensuring amplitude variety in the output.
- Calculated note duration, velocity, and if the note was repeated from the previous step.
- Ensured that a maximum of n notes were active at one time, removing notes with the lowest velocity. This was a further control for overall density.
- Combined repeated notes across n steps. This avoided the restatement of notes at every time step, but allowed restatement according to a threshold.

21. Chen, 'Voice Without Speaker: human speech synthesis in acoustic instrumental contexts', 67.

22. Rachel M. Bittner et al., 'A Lightweight Instrument-Agnostic Model for Polyphonic Note Transcription and Multipitch Estimation', in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (Singapore, 2022), <https://doi.org/10.1109/ICASSP43922.2022.9746549>.

- Mapped Mel-bands to the corresponding MIDI notes to control the motors, and converted the energy value of each band to a MIDI velocity.

a. Code/Chapter 4 - Motorcase II code/Transcription.py

4.3.6 Improvisation

To develop an improvisational practice based on navigating the dataset's emotional labels, I designed a software interface around a simple MIDI controller. This was structured around two modes of interaction: navigating the emotional space of the corpus and manipulating the transcribed material in real-time. Using an existing controller meant I could easily add or modify controls in response to the experience of playing the device.

Primary navigation controls were intended to explore the three-dimensional emotional space of the MSP-Podcast dataset. My goal was to make this vast corpus intuitively playable, and while there was no visual feedback, the physical sliders of the MIDI controller showed the approximate position in the dataset. Once a file was selected, a limited set of controls provided tools to manipulate it, through looping, pausing, and remapping.

Code Example 4.6 — *Motorcase II*: Interface Detail

Navigation and manipulation controls for the Max interface^a are described below.

Navigation controls:

- **Sliders 1–3: VAD Adjustment**
Mapped logarithmically to ensure even representation of files across the range, compensating for the normal distribution of metadata.
- **Slider 4: Nth Closest**
Expands the selection at any coordinate by returning the *nth closest* file, overcoming MIDI slider resolution limits.

Manipulation controls:

- **Slider 5: Loop length**
Sets the playback duration from 0.1 seconds to the full file length.
- **Button A: Play/Pause**
Pauses playback and 'freezes' motors in their current state.
- **Button B: Remap**
Randomises MIDI mapping to motors, allowing varied realisations of the same data.

a. Code/Chapter 4 - Motorcase II code/Improvisation.py

Learning to play *Motorcase II* required developing a familiarity with its output. While I found the breadth of the dataset made individual files impossible to anticipate, a gradual learning of the VAD regions allowed for a degree of intentionality. The ambiguous middle regions of the dataset often led to surprising results, reinforcing my role in performance as shaping output rather than through direct control. Repetition introduced another layer of unpredictability: due to variations in actuator starting positions, the same MIDI file would sound different each time it was played. This physical variability was an expressive feature, allowing for subtle changes within repeated loops.

In live improvisation I often alternated between three modes: (1) short loops where physical variations provided interest, (2) sweeping through dataset regions while searching for new material, and (3) manipulating found material through gradual changes to loop length or sudden timbral shifts using the remap function. As with *Motorcase I*, the link between the output and its speech source was hidden.

Portfolio Example 4.6 — *Motorcase II*: Improvisation Examples

A solo video shows some of the approach to improvisation:^a ►

- 0:00—01:27 uses a single sample through adjustment of loop length.
- 01:09 shows the unpredictability of the objects and actuators creating variation.
- 01:46 shows the freezing of a sample to continue a motor during other samples.
- 04:37 shows the remapping function, as the same loop is mapped to a new set of motors.
- 05:40 shows the general approach of searching for material to loop and adjust.

A duo EP was recorded with Gaia Blandina. Here, the interaction involved reacting to the device's unpredictable output:^b

- Track 1 focuses on the freezing control, allowing single motors to spin for extended periods. ►
- Track 2 begins with scrubbing through the dataset, to find material to use. 01:18 involves playing inside the case with the fingernails, with motors activated again at 02:25. ►

- Track 3 adjusts the loop length of a sample, remapping to different objects in the case, and returning to a short loop at 02:10. ►
- Track 4 combines manual playing in the case, and held motors. 01:10 moves through the dataset to alight on different samples for short periods. ►

a. Portfolio/Chapter 4 - Motorcase II/Solo improvisation.mov
 b. Portfolio/Chapter 4 - Motorcase II/Duo EP/...

4.3.7 Composition

To apply the system in combination with instruments and electronics, I wrote *Machine Translation* (2025) for *Motorcase II*, bass clarinet, violin, drumkit, and a tape part featuring RAVE²³ generated speech. Where improvisation with the device focused on navigating the corpus, this fixed piece allowed detailed musical exploration of specific VAD regions, and the integration of generated audio.

Structured in four movements, each corresponded to a different VAD region of the dataset. I initially generated the material for each movement through improvisation with the device, recording both MIDI and audio. This initial material was intuitively shaped and refined. The resulting motor part was then triggered from a MIDI keyboard rather than the improvisational interface.

RAVE models were trained on VAD regions of the speech dataset. Instrumental parts derived from the motor transcriptions were then used as the input for these models. I intended this to create a loop in the compositional process: speech was transcribed to the motors, again transcribed and performed by the musical instruments, and finally re-synthesised into generated speech by the RAVE models.

Model Example 4.1 — *Motorcase II*: RAVE models

Each RAVE model and training parameters are made available.^a

- 1_1_1_v3_caeb149cc6_streaming.ts dataset region 1, 1, 1 (VAD). 
- 1_3_7_v3_caeb149cc6_streaming.ts dataset region 1, 3, 7. 
- 2_4_6_v3_caeb149cc6_streaming.ts dataset region 2, 4, 6. 

23. Caillon and Esling, 'RAVE: A variational autoencoder for fast and high-quality neural audio synthesis'.

- 7_7_7_v3_caeb149cc6_streaming.ts dataset region 7, 7, 7. 

a. Models/Chapter 4 - Motorcase II models/...

My aim with the piece was to blend the three elements described: the sound of the device, acoustic instruments, and generated speech from the RAVE models. Generated audio was used sparingly, often at a low dynamic level, to add vocal colour and emphasise specific rhythmic or gestural aspects of the instrumental lines. As with the composition for *Motorcase I*, instrumental techniques were chosen to blend with the sonic qualities of the device.

Portfolio Example 4.7 — *Motorcase II: Composition Examples*

Aspects of the compositional approach are highlighted below, relating to the recording^a ► and score:^b ≡

- Movement 1 takes fragments of the dataset region VAD 1,1,1 as the starting point for repeated musical gestures, recombining these as the movement progresses.
- Movement 2 from 02:07 takes a similar approach to dataset region 1,3,7, with rhythmic loops. The RAVE model can be heard at 03:19, adding vocal colour.
- Movement 3 from 04:09 takes as a starting point an improvisation in region 2,4,6 which a single loop is remapped to the motors, harmonised by the instruments as the device gradually deviates from the loop.
- Movement 4 from 06:23 takes material from region 7,7,7 in alternating rhythmic loops and single repeated notes, with the RAVE models clearly heard at 06:47 though mainly used to emphasise percussive aspects, as in 07:52.

a. Portfolio/Chapter 4 - Motorcase II/Machine Translation.wav

b. Portfolio/Chapter 4 - Motorcase II/Machine Translation - score.pdf

4.3.8 Evaluation

This evaluation follows a parallel structure to that of *Motorcase I*, following the sequence of translation from the corpus of speech recordings to the electromechanical output of the device. Each stage contributes to the overall goal of this chapter: to render speech through motors and found objects in an improvisational and compositional practice. However, the project's particular focus on emotional speech qualities is also addressed.

Source to Transcription

The vast MSP-podcast corpus provided a wide variety of speech-derived musical output. The transcription process itself was, like *Motorcase I*, reductive—by grouping frequencies into thirty-two bands and then selecting only the top five most active bands, the audio was quantised according to pitch, and simplified. This can be seen in Figure 4.12, where a resulting MIDI transcription is overlaid on a spectrogram of speech audio. Despite this reduction, a high degree of rhythmic correspondence is evident when speech audio and motor output are played simultaneously (Portfolio Example 4.8).

Portfolio Example 4.8 — *Motorcase II*: Transcription Example

Overlaying audio and motors together shows an overall rhythmic correspondence.^a ► Finding a balance between imitation and novelty was an iterative process of fine-tuning the transcription parameters, relying on an intuitive judgement of the resulting material.

^a. Portfolio/Chapter 4 - Motorcase II/Examples/Transcription example.wav

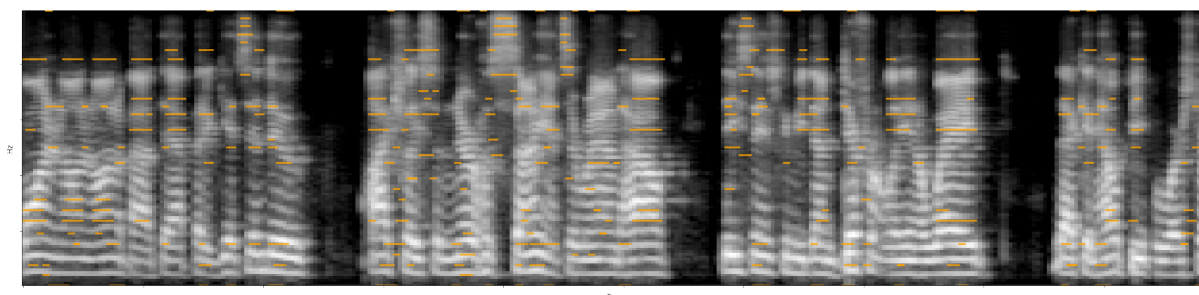


Figure 4.12: *Motorcase II* MIDI transcription (orange) overlaid onto a spectrogram of the original audio (grey), showing the relationship between the transcribed data and speech source.

Given the reduction inherent to the transcription process, a question for the project was to what extent the musical output would align with the emotional labels of the source material. Through improvisation, I observed that the output varied with VAD values, particularly at the extremes of the scale: the region associated with anger (low valence, high activation and dominance) produced denser and more irregular rhythmic patterns, whereas the sadness-related region featured longer motor durations and lower amplitude variation. These subjective observations align with existing research on emotional speech prosody.²⁴

²⁴. Juslin and Laukka, 'Communication of emotions in vocal expression and music performance: Different channels, same code?'

To evaluate these observations, I used heat-maps to visualise changes in broad rhythmic and dynamic features across the transcribed MIDI in VAD space. Features chosen for analysis included average onsets per second (Figure 4.13), note duration range (Figure 4.14), average velocity, and silence ratio. The results for onset rates broadly corresponded with existing research: regions associated with anger had the fastest speaking rate, while sadness exhibited slower rates. However, some differences were present: MIDI velocity did not align with expected intensity differences, likely due to variations in the dataset's recording conditions. This, I accepted as part of the process.

While these findings were limited in scope, I found them highly useful for performance. They revealed a sonic topography of the device, such as higher activation and dominance values correlating with increased onset density. This data, alongside intuitive observations about the device's properties, allowed a more intentional approach to improvisation.

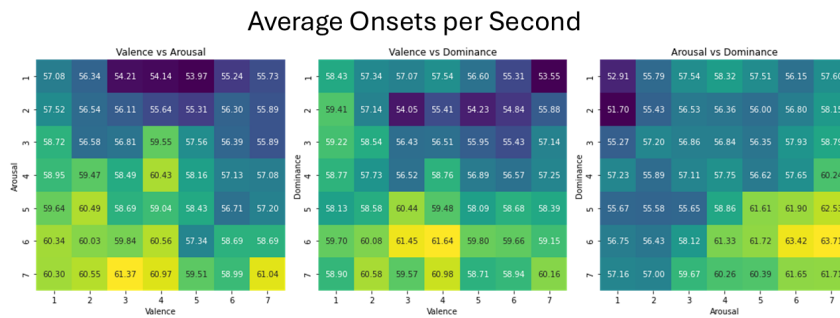


Figure 4.13: Heatmap showing changes to average onsets per second across the VAD space of the MIDI transcriptions

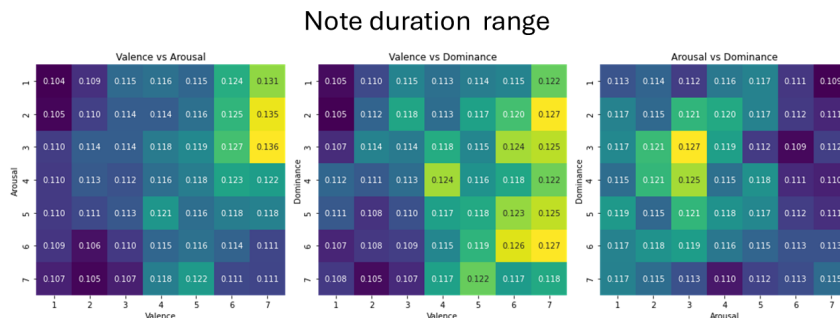


Figure 4.14: Heatmap showing changes to note duration range across the VAD space of the MIDI transcriptions

Transcription to Device

As with *Motorcase I*, the process of translating speech into mechanical sound introduced a series of discrepancies between the transcription and the motor realisation. The thirty-two servos of the device essentially create a physical, acoustic version of a Mel-spectrogram. However, the objects were not arranged deliberately to recreate the width of the original spectrogram. Instead, each object represented a Mel-band, activated when that band of the original speech was present. This decoupled the spectral qualities of the voice from pitch layout, instead distributing the bands of the analysis across the varied resonance of the metal objects.

The physical instabilities of the device added a layer of instability to the realisation. The actuators, from the wobble of a spring to the bounce of a bead, meant that identical MIDI data never produced an identical sound—an example of this is shown in Portfolio Example 4.6 (01:09).

Improvisation and Interface

The interface was designed primarily for the navigation of the corpus, with the device responding to a given VAD-region with the nearest MIDI transcription. In mapping the three-dimensional VAD space to the controller, this allowed for the improvisational exploration of the dataset's range. Here, the MIDI controller was a navigational tool for the topography identified through VAD analysis.

The manipulation controls, of loop length, remapping, pausing and freezing were essential tools in performance for creating musical interest, and in a duo setting allowed a degree of real-time control over the output of the device. However, because the scale of the corpus meant total familiarity with the individual recordings was impossible, I found the performance process primarily reactive. This involved a cycle of triggering a VAD region and then using the manipulation controls to shape the resulting mechanical output in the moment.

Composition

The design of the device around the navigation of a large-scale corpus created a tension with a requirement for repeatability in a fixed composition—returning to specific, previously encountered transcriptions was usually impossible. To bridge this, the compositional process for *Machine Translation* involved recording the acoustic and MIDI output of improvisations and subsequently

choosing fixed sections for the ensemble and tape. This workaround produced novel material for the composition, but suggests an incompatibility between the device's improvisatory design and my process of writing a fixed composition. This is not to say I wanted absolute repeatability, as the inconsistencies of the mechanical output were an important part of the piece, but the near total inability to return to previously heard material was a challenge in composing with the device.

The broader conceptual loop between motors, instruments, and RAVE resynthesis provided a frame for the work. Using the RAVE model was an attempt to close the gap between the speech source and consequent mechanical sounds of the device, in a way that hinted at the speech source rather than relying on audio recordings of the original speech. I found this approach allowed the blending of different elements in an ensemble setting: the motor realisation, instrumental parts, and resynthesised speech.

4.3.9 Summary

Motorcase II scaled up the practice established in the first project, shifting from a small, manually-assisted dataset to a vast corpus, and towards a statistical and topographical approach to speech-derived material. I explored a novel method for navigating a large-scale dataset through emotional labels, which resulted in an automated transcription method enacted through a self-contained device. Evaluation of the musical output suggested correspondence between the VAD labels and the sonic character of the transcriptions, though more work would be needed for a thorough assessment.

While the scale of the corpus made returning to specific transcriptions difficult, this was a consequence of the device's design for real-time navigation. Overall, the project demonstrated that a vast dataset can be rendered as a coherent and performable mechanical instrument, where the mapping of emotional labels provides the method for its navigation.

4.4 Discussion

This section addresses R1, on how speech features can be computationally extracted and implemented in a mechanical improvisation and composition practice. This is structured in three parts, drawn from the research question and chapter objectives: methods of abstraction, material

realisation, interaction and performance.

4.4.1 Methods of Abstraction

Transcribing and performing speech audio through each device erased linguistic content to focus on the approximate transcription of speech features, a process which draws on Miller’s concept of transformative mimicry. The first stage of transformation was algorithmic. In *Motorcase I*, the abstraction was based on the segmentation of speech, in a manually-assisted method afforded by the small-scale dataset. In *Motorcase II*, the abstraction was a reductive analysis via the Mel-spectrogram, required for the automated processing of a large-scale corpus (Table 4.2).

Table 4.2: Abstraction and Effect in *Motorcase I* and *II*

	<i>Motorcase I</i>	<i>Motorcase II</i>
Features	Text, amplitude, spectral centroid.	32-band Mel-spectrogram, VAD labels.
Abstraction	Segmentation based on text.	Spectral reduction and peak-picking.
Effect	Letter segmentation $\Rightarrow$ expanding to multi-onset rhythms Word segmentation $\Rightarrow$ reduction to single-onsets Phrase segmentation $\Rightarrow$ flattening into single events	Spectral reduction $\Rightarrow$ quantised frequency Peak picking $\Rightarrow$ isolation of dominant spectral peaks

Both *Motorcase* devices sit between a transformative and generative mapping, with performer control tilting the devices away from an unmediated presentation of the speech transcriptions. However, the decisions I made in the transcription process defined the output of the devices, and align with Cassandra Miller’s reflection that such transcriptions ultimately reveal more about ‘choices regarding the transformation... than about the source itself’.²⁵ The choice of abstraction filters the source speech; while a causal link to the original audio contour and rhythm is preserved, the musical features of the transcription are products of the specific method.

4.4.2 Material Realisation

The physical enactment of the transcription data represents a second stage of transformation. In this stage, the abstracted features of the transcription data are mediated by the mechanical properties of the motors and the acoustic properties of the found objects.

25. Miller, ‘Transformative Mimicry: Composition as Embodied Practice in Recent Works’.

The found objects provide the broad acoustic character for each device. Figure 4.15 shows the Long-Term Average Spectrum (LTAS) for both devices, illustrating how the material palette defines the sound world. The differences represent the material qualities of each project. *Motorcase I* used a mix of plastic, wood, glass and metal, while *Motorcase II* was restricted to an all-metal palette.

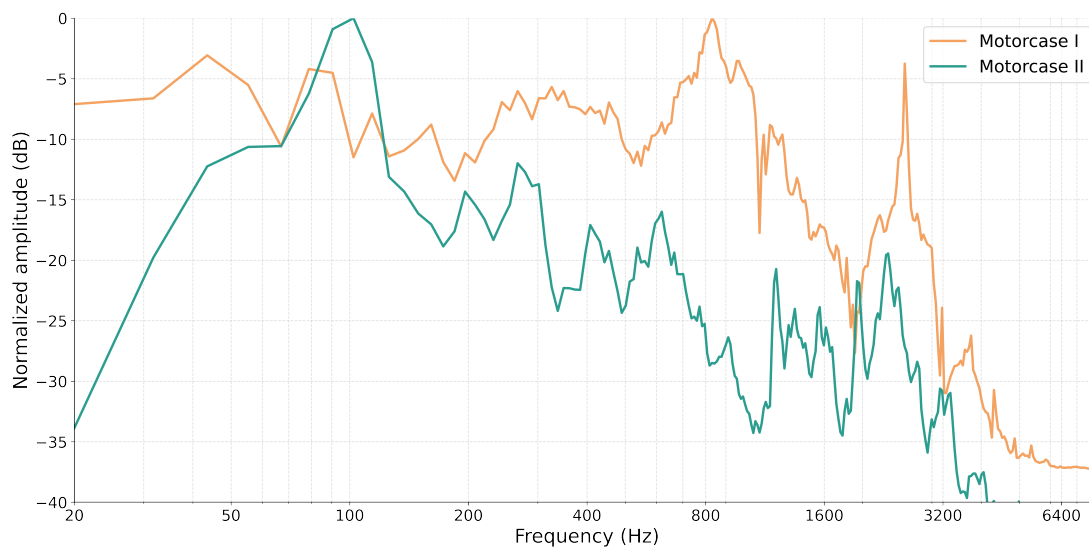


Figure 4.15: Normalised LTAS of *Motorcase I* and *II*. *Motorcase I* (orange) exhibits a broader distribution resulting from its mixed material palette, while *Motorcase II* (green) shows more distinct peaks with its all-metal construction.

Differences between the transcription and mechanical realisation are primarily a result of the physical properties of the found materials, and the configuration of each object-actuator. As the resulting material is a combination of the source and realisation mechanism, to what extent is it influenced by either? Ablinger uses a visual metaphor—of photographic resolution—to describe the *Quadraturan* process. In the case of the *Motorcase* devices, I find a *mosaic* to be a useful analogy; each mechanism has a particular internal pattern, which is superimposed on the form of the speech transcription.

Take two mechanisms, the first (Figure 4.16) a rotating wheel which sounds in a gradual rise-and-fall pattern. A second (Figure 4.17) has two objects struck by a rotating arm, producing two distinct attacks. Superimposing these patterns onto a transcription (Figure 4.18)—represented here as a two-event sequence—will produce new musical material with the form of the transcription, and the surface pattern of the mechanisms. The result is a balance of the two, and the degree to which it tilts towards the source or realisation mechanism is determined by the qual-

ities of the mechanism, and the resolution of the transcription: a higher resolution transcription will reflect the source form in greater detail, while an unpredictable mechanism will disrupt this representation of the form.

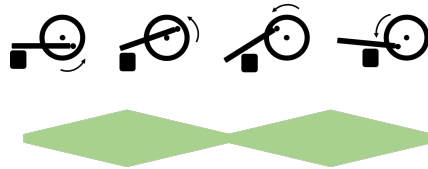


Figure 4.16: Mechanism 1, a rotating wheel which produces a rising-falling sound (green).

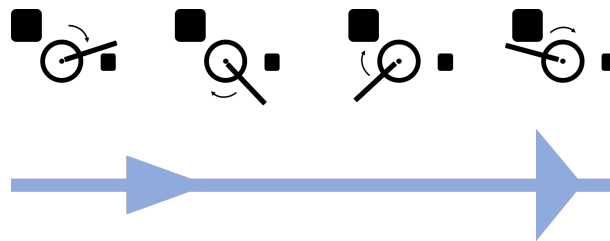


Figure 4.17: Mechanism 2, a rotating arm striking two objects to produce two distinct attacks (blue).

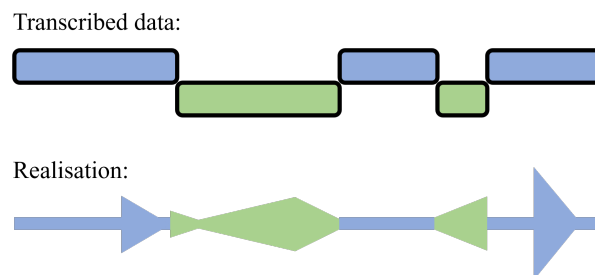


Figure 4.18: The realisation, in which the sound of each mechanism is superimposed on the form of the transcribed data. The transcription here is represented simply as a sequence of two events, mapped directly to each mechanism.

This is the transformative process, in which the qualities of the objects define the surface while the form is ultimately derived from the speech recording. Compounding these differences are mechanical limitations, including the speed at which the motors could respond to instructions. In the *Motorcase* devices, I use these mechanical inconsistencies to generate variation within looping sequences of the transcriptions, a controlled unpredictability emerging from the system's material qualities.

4.4.3 Interaction and Performance

The final stage of transformation is performative, involving the real-time manipulation of the transcription data. In both projects, I felt that a literal rendering of the speech features was insufficient for live performance; the mapping required extension through algorithmic and stochastic procedures to create variation. This strategy is found in other existing works, for example Valle's *Rumentarium* where binary mappings were varied through motor voltage and text-scanning speed,²⁶ and Dunham's *Click::TWEET*, which uses spatialisation and rhythmic disruption to expand a Morse code mapping.²⁷

In the *Motorcase* works, these manipulations were navigated through two categories of control: those derived from the transcription features themselves, and performative controls based on broader electro-acoustic techniques (Table 4.3). The improvisational language that emerged was a result of balancing playback with performative intervention. Parkinson and Dunning describe this as a machine that 'plays' rather than 'runs', where unpredictability allows the specific mechanical behaviour of the system to materialise in the sound.²⁸

Table 4.3: Performance Controls for *Motorcase I* and *II*

	<i>Motorcase I</i>	<i>Motorcase II</i>
Transcription-linked	Segmentation resolution	Regional selection via emotional labels Remapping transcription to motors
Performative	Direct sample selection Mixing samples Loop length Speed control Adjusting proportion of silence Quantisation Individual motor control	Loop length Pausing / Freezing

As Table 4.3 shows, the transition from *Motorcase I* to *II* involved a significant reduction in performative controls. I would explain this in part by the scale of the dataset, with the performative controls of *Motorcase I* a response to the aesthetic concern of generating enough variation to sustain a performance. For me, the vast dataset of *Motorcase II* provided inherent variety, allowing my focus to shift away from manual manipulation and towards the navigation

26. Valle, 'Making acoustic computer music: The Rumentarium project'.

27. Paul Dunham, 'Click:: TWEET. A Telegraphic Sound Installation', in *Proceedings of the International Conference on New Interfaces for Musical Expression* (2021), <https://doi.org/10.21428/92fbeb44.48c39866>.

28. Adam Parkinson and Graham Dunning, 'Improvising with Machines: A taxonomy of musical interactions', *Organised Sound* 29, no. 3 (2024): 340–351, <https://doi.org/10.1017/s1355771824000268>.

of the corpus.

In both, I would suggest the theatre of the mechanical objects provided a visual and narrative layer to the performance, and a humour emerged in both works through the coordinated electromechanical movement and the constant threat of the system falling apart. I interpret this as a tension created by the absurdity of complex computational data driving unstable found objects, between the precision of the software and the precariousness of the hardware. Here, the theatricality of the works seemed to reinforce the controlled unpredictability described in the previous section; the performance becomes a balancing act between an intended form and the physical instability of the motors.

While the original speech remained a hidden driver during improvisation, in the compositions for each project I selectively reveal the source. In *Years a have to know or something*, I made this link explicit through the use of archival video, providing a direct reference for the mechanical realisations. In *Machine Translation*, I used RAVE-generated speech to provide a subtle timbral enhancement, to shadow the mechanical output. In both, the aim was to make a causal link between the speech source and the mechanical sound world more transparent to the audience.

4.4.4 Conclusion

This chapter has presented two creative applications of algorithmic speech transcription in a DIY composition and improvisation practice. To address R1, I have explored the extraction of features including text-based segments, spectral centroids, amplitude, and Mel-spectrograms, and their implementation within three stages of musical transformation: through algorithmic transcription, material objects, and performative control. I found that the implementation of speech features in this practice involved a balance, between the linguistic form of the source and the material qualities of the found objects and actuators. Within this approach, I demonstrated two approaches to generating musical variety: *Motorcase I* centred on the manipulation of a few known sources, whereas *II* was based on the exploratory navigation of a vast corpus.

The findings and limitations of the *Motorcase* projects directly inform the subsequent research. The use of generated speech audio with RAVE, explored briefly here, becomes the focus of the following chapter. To further develop the abstraction of speech, Chapter 6 introduces text models as a method for generating acoustic-mechanical renderings of syntax patterns.

Chapter 5

Audio Models in Practice

5.1 Introduction

In this chapter I incorporate NAS generation into the DIY motor-driven practice developed in Chapter 4, through two works: *Smalltalk*, and *Radio User Guide*.

While the previous chapter focused on the transcription and mechanical realisation of speech recordings, in the works here I use neural synthesis to generate the primary audio material. RAVE¹ was chosen for this purpose. Unlike Text-to-Speech or concatenative synthesis, RAVE is a general-purpose model that learns to represent and reconstruct the raw audio signal. By training a model on speech datasets, audio can be generated with the timbral qualities of the voice while remaining decoupled from linguistic meaning—a feature central to the exploration of asemic and fragmented speech throughout this chapter.

Chapter 2 revealed synthesised speech as a source in several existing works, from the LPC synthesis of DeMarinis’ *Alien Voices*² to the more recent neural synthesis of Walshe and Akten’s *ULTRACHUNK*.³ However, Chapter 3 showed few examples of the integration of generated speech into electromechanical practice. In this chapter, my aim with both works was to explore the use of speech-trained audio models in combination with electromechanical sound, guided by R2:

1. Antoine Caillon and Philippe Esling, ‘RAVE: A variational autoencoder for fast and high-quality neural audio synthesis’, *CoRR*, 2021, <https://doi.org/10.48550/arXiv.2111.05011>.

2. Paul DeMarinis, *Alien Voices*, 1988, accessed 25 February 2026, <https://pauldemarinis.org/AlienVoices.html>.

3. Memo Akten and Jennifer Walshe, *ULTRACHUNK*, 2018, accessed 25 February 2026, <https://memo.tv/works/ultrachunk/>.

R2. How can NAS models trained on speech datasets be incorporated into this practice, and what methods can be used to organise generated material?

To address this, both *Smalltalk*, and *Radio User Guide* involved training RAVE models to produce a high volume of speech-derived material, raising the question of how to organise these outputs compositionally. This chapter concerns both the practical incorporation of these models into the artistic practice, and the methods used to identify and structure musically useful material.

My aesthetic approach is informed by Privato and Magnusson’s concept of *AI Hauntography*, which emphasises the technological uncanny emerging from fragmented and decontextualised outputs.⁴ This uncanny quality is heightened when working with the voice; consequently, this chapter focuses on the unpolished characteristics of models, searching for unexpected articulations, failures, and instabilities.

As shown in Table 5.1, the two works present contrasting approaches to intelligibility and organisation. In *Smalltalk* I draw on the concept of asemic writing to organise material around pitch and amplitude contours. *Radio User Guide* focuses on semantic fragmentation, moving from poorly trained, unintelligible models toward gradual coherence. In both, I use groups of three devices to place the machines in an apparent dialogue, though the final works remain fixed compositions.

Table 5.1: Overview of Project Elements: *Smalltalk* and *Radio User Guide*

	Smalltalk	Radio User Guide
Dataset	Audiobooks, whispers, overlapping speech	Audiobooks, online dictionaries, music
Approach to Synthesis	Timbral transfer	Iterative training to capture the emergence of speech
Organisational Method	Asemic, matching similar pitch and amplitude contours	Semantic fragmentation and gradual coherence
Devices	3 devices: motors, springs, metal plates, speakers	3 devices: motors, found radio parts, speakers
Creative Output	Fixed composition (devices)	Fixed composition (devices, human performers)

4. Nicola Privato and Thor Magnusson, ‘Querying the Ghost: AI Hauntography in NIME’, in *Proceedings of the International Conference on New Interfaces for Musical Expression* (2024), 432–438, <https://doi.org/10.5281/zenodo.13904901>.

Objectives

There were three key objectives for the portfolio works discussed in this chapter:

1. **Speech Synthesis:** Train RAVE models on varied speech datasets, generating material that emphasises unstable and unexpected characteristics.
2. **Organisational Method:** Develop strategies for selecting and structuring generated material into fixed compositions.
3. **Devices:** Construct three motor-driven devices for each work, using speakers for audio playback and a simplified material palette of found objects.

5.2 Smalltalk

5.2.1 Overview

Smalltalk is a composition performed by three box-like devices, each with a speaker and two motors. In Chapter 2, I pointed to ‘deliberately obscuring semantic meaning’ as a shared artistic intent, and is of particular relevance to this creative project where my approach was influenced by asemic writing: a visual art practice where the graphical qualities of letterforms are detached from semantic meaning.⁵ This focus on the shape of expression is an analogy for engaging with the speech-like, but non-semantic output of the RAVE models.

In Chapter 3, electromechanical works taking speech as a source often focused on semantic meaning—whether clearly articulated through a large language model in Zuiderveld’s *Conversations beyond the ordinary*,⁶ or transcribed from audience speech to be fragmented and obscured in Bacon’s *PROBE II, Subaudition*.⁷ In the project presented here, rather than treating text or speech as a source to be fragmented, the audio generated by speech-trained RAVE models is already stripped of semantic meaning.

In this context, one question is how to organise this generated material compositionally. To write the work, I explore a novel approach to selecting and composing with audio generated by RAVE models, combining computational analysis with intuitive decision-making via a drawing

5. Peter Schwenger, *Asemic: The art of writing* (University of Minnesota Press, 2019).

6. J. Zuiderveld, *Conversations beyond the ordinary*, 2024, accessed 25 February 2026, <https://ars.electronica.art/hope/en/conversations-beyond-the-ordinary/>.

7. Benjamin Bacon, *PROBE II, Subaudition*, 2021, accessed 25 February 2026, <https://www.benjaminbacon.studio/probe-1>.

interface. I use this simple tool to sift through generated audio by drawing shapes, using a dynamic time warping (DTW) algorithm to pick out similar pitch or amplitude contours. The audio resulting from this selection process was then organised manually into patterns and sequences to highlight the characteristics of the generated material.

Objectives

1. **Dataset:** Curate audio of ordinary, whispered and overlapping speech to provide the basis for RAVE model training.
2. **Synthesis:** Train several RAVE models to produce a variety of speech-derived audio.
3. **Organisation:** Develop a drawing interface to identify and select segments based on pitch and amplitude contours.
4. **Device:** Construct three identical devices using motors and springs to support apparent dialogue.
5. **Composition:** Arrange the selected materials into a fixed work for the three devices.

5.2.2 Dataset and Training

The goal was to create a broad palette of speech-derived sound, with a range of phonetic sounds, pitch and amplitude contours, and rhythmic features. To provide this varied basis for the work, I selected between four and six hours of audio from two distinct speech corpus:

1. Whispered speech from the CHAINS speech corpus.⁸
2. Ordinary speech from the Librispeech dataset.⁹
3. Overlapping speech created by layering audio from Librispeech.

Training took between eight and ten days for each RAVE model. I chose the training configuration through trial-and-error, based on an intuitive assessment of the output and how this fit with the aim of the piece: the whispered speech and overlapping speech models used ‘v2’

8. Fred Cummins et al., ‘The chains speech corpus: Characterizing individual speakers’, in *Proc of SPECOM* (2006), 1–6, <https://doi.org/10.35111/cgbv-ke56>.

9. Vassil Panayotov et al., ‘Librispeech: an asr corpus based on public domain audio books’, in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (IEEE, 2015), 5206–5210.

without augmentation¹⁰. The only difference for the ordinary speech model was the use of the ‘wasserstein’ configuration—it has been argued this can result in higher quality synthesis.¹¹

5.2.3 Synthesis

Once trained, I used these models to generate audio through two methods. First, external audio was passed through each model for timbral transfer,¹² using a variety of source samples including percussion, violin, speech, noise, and found objects.

Second, Valenzuela’s RAVE-Latent-Diffusion model¹³ was used for unconditional generation, to produce material directly from the models’ latent space.¹⁴ By adjusting the temperature parameter during generation, the degree of variation and randomness in the output could be controlled; reducing the temperature focuses the model on the most probable latent trajectories, while increasing it allows for a wider, more random exploration of the latent space.

The overall output of these models was a large and varied collection of non-semantic, speech-like audio. Recognisable consonants and vowels appeared in the output of all models, but in my estimation the RAVE-Latent-Diffusion model proved the most effective at producing convincing speech-like qualities.

Model Example 5.1 — *Smalltalk*: Models

Each trained model is made available.^a ▢ The use of a particular training configuration was the result of an intuitive evaluation of the model outputs.

^a. Models/Chapter 5 - Smalltalk models/...

10. The dataset was large enough to avoid augmentation here, and testing various augmentations—in my listening—did not improve the model. For alternative configurations see Antoine Caillon and Philippe Esling, *RAVE: Official implementation of the RAVE model*, 2021, accessed 25 February 2026, <https://github.com/acids-ircam/RAVE>

11. Di Hu and Katunobu Ito, ‘Two-stage instrument timbre transfer method using RAVE’, in *2024 International Symposium on Multimedia (ISM)* (2024), 33–40, <https://doi.org/10.1109/ISM63611.2024.00011>.

12. The timbre transfer technique maps the qualities of an input sound onto the timbral characteristics represented by a trained RAVE model. For more information see Caillon and Esling, ‘RAVE: A variational autoencoder for fast and high-quality neural audio synthesis’, 15

13. Moisés Horta Valenzuela, *Rave Latent Diffusion*, 2023, accessed 25 February 2026, <https://github.com/moiseshorta/RAVE-Latent-Diffusion>.

14. While standard RAVE models require an external audio signal to provide the latent data for synthesis, RAVE-Latent-Diffusion uses a diffusion process to generate this data from scratch—see Horta Valenzuela

5.2.4 Organisation

To address the challenge of selecting musically useful material from the collection, I used a drawing interface and a Dynamic Time Warping (DTW) algorithm¹⁵ to identify segments based on pitch and amplitude contours. My purpose here was not identifying exact contours, but finding micro-gestures within the generated audio that followed particular trajectories.

The steps for this selection process were as follows:

1. An audio file is selected to analyse.
2. The user draws a shape within an input box on the interface.
3. The script compares the drawn shape to the contour of either the fundamental frequency or the amplitude of the audio signals.
4. The interface presents a visual plot of the original drawing alongside the contours of the best-matching audio segments.
5. An interactive list of the top matches as playable audio extracts is returned, ordered from best to worst match.

This tool allowed for the selection and evaluation of material based on approximations of the fundamental frequency or amplitude contour (Figure 5.1). The final step was intuitive: from the list of matches provided by the algorithm, I made a selection, saving the results to a folder to form a working palette.

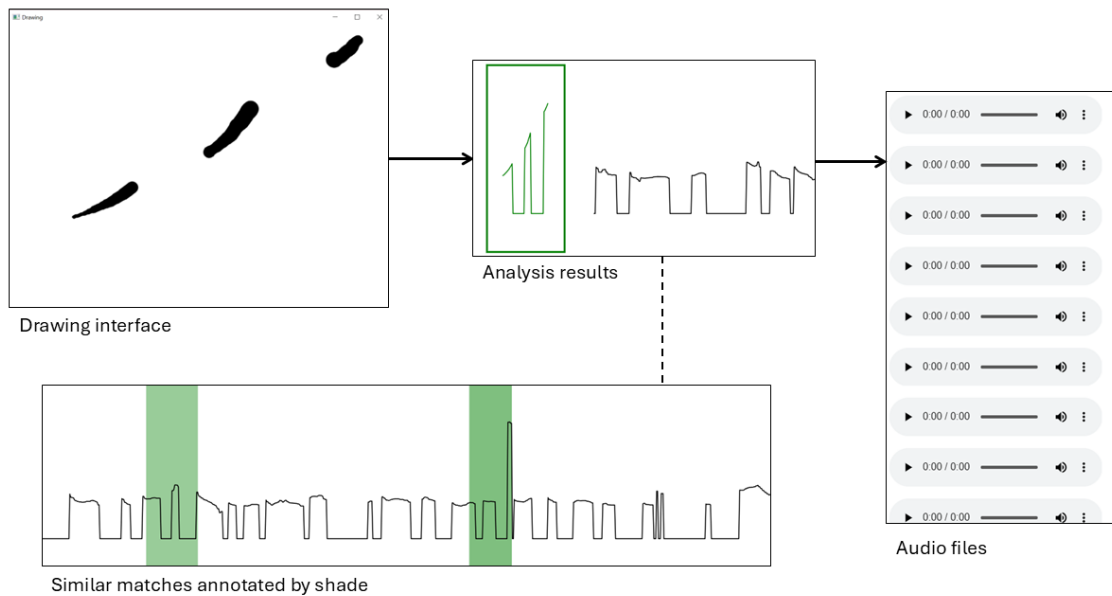
Code Example 5.1 — *Smalltalk*: Contours

A Python script^a `</>` was run in an interactive notebook to extract material with similar contours. A second script^b `</>` was used for longer extracts, where rendering the image in notebook was not possible, and was saved instead.

^a. Code/Chapter 5 - Smalltalk code/MatchingGestures.py

^b. Code/Chapter 5 - Smalltalk code/MatchingGesturesLongplot.py

15. DTW is an algorithm used to measure the similarity between two temporal sequences that may vary in speed or timing. By warping the time axis of the sequences, DTW calculates an optimal match between their shapes, making it effective for identifying similar sequences regardless of their duration. See Meinard Müller, ‘Dynamic Time Warping’, in *Information Retrieval for Music and Motion* (Berlin, Heidelberg: Springer Berlin Heidelberg, 2007), 69–84, https://doi.org/10.1007/978-3-540-74048-3_4

Figure 5.1: *Smalltalk* drawing interface

5.2.5 Device

I intended the work to be staged as a three-way conversation, between three separate devices (Figure 5.2). The basic function of each was to play audio and produce acoustic sound via two motors, springs, and a metal plate.

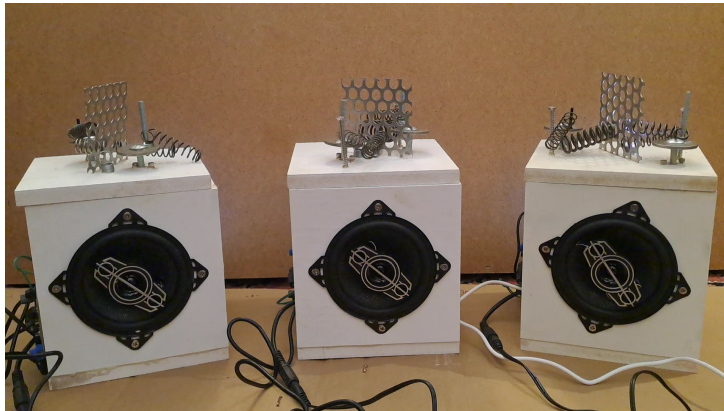
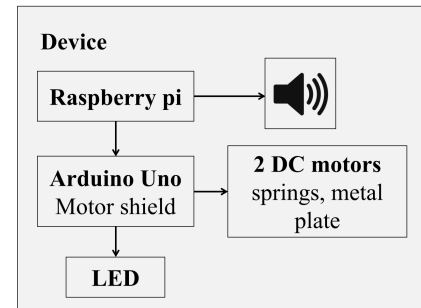
For each device, I used a Raspberry Pi as the central controller, running a Python script for audio and motor playback. Audio was output via a small amplifier and single speaker. The Pi communicated via serial messages with an Arduino, which controlled a single LED and two DC motors. Each motor was fitted with a spring, which in turn actuated a second spring attached to a small, perforated metal plate (Figure 5.3).

Code Example 5.2 — *Smalltalk*: Communication

One device ran a script^a reading a MIDI file containing the composition. It played its own part while sending instructions for the other two devices over OSC. The two receiving devices ran a corresponding script.^b

^a. Code/Chapter 5 - Smalltalk code/SmalltalkSEND.py

^b. Code/Chapter 5 - Smalltalk code/SmalltalkRECEIVE.py

Figure 5.2: *Smalltalk* devicesFigure 5.3: *Smalltalk* design

5.2.6 Composition

The final composition, *Smalltalk*, was assembled by arranging the selected audio palette into a structured work for the three devices. I followed three primary methods of grouping and sequencing material:

1. Raw Fragments: Some unmodified excerpts from the RAVE models were used directly.
2. Contour Groups: The majority of the composition consists of fragments selected via the drawing tool. These were arranged into sequences and patterns that explored a particular shape or direction, creating rhythmic and melodic phrases from shared characteristics.
3. Phonetic Groups: Some fragments were chosen manually based on their phonetic qualities (e.g. collections of specific vowel or consonant groups, identified by ear).

The instructions for the motors were added as a parallel layer via manual transcription. By listening to the arranged audio and transcribing the rhythms into MIDI commands, the motors and springs were made to shadow the audio playback. I used this manual approach to ensure a basic synchronisation between the electronic sound and the physical movement of the motors.

The resulting performance uses the spatial arrangement of the devices to suggest a conversation, in fragmented, non-semantic speech. While the work functions as an experiment in pairing non-semantic neural synthesis with mechanical sound, the relationship between the two remains primarily one of shadowing: the metallic scraping and clanging of the springs follows the contours of the RAVE audio as a physical accompaniment.

Portfolio Example 5.1 — *Smalltalk*: Video

Extracts of the recorded work^a ► are highlighted below:

- 0:00 raw material from the whispering model, layered with motors.
- 02:13 sequences of repeating gestures from the whispering model.
- 04:17 raw material from the speech model.
- 05:13 fragments extracted by pitch shape, arranged into repeating patterns.
- 06:02 speech fragments in patterned sequences.
- 07:37 raw material from the crowd model.
- 07:57 crowd model output gradually organised into phonetic groups.

^a. Portfolio/Chapter 5 - Smalltalk/SmalltalkVideo.mov

5.2.7 Evaluation

The following evaluation follows the process of synthesis, organisation, and mechanical implementation, from the initial training of RAVE models to the realisation of the composition.

Dataset to Synthesis

In the context of this piece, I used RAVE models as a functional method for generating a palette of non-semantic speech-like audio. As is typical when training these models on speech data, the synthesis captured the spectral characteristics of the source but unconditional generation did not maintain the temporal coherence required for intelligible words. These articulatory qualities, such as the breathiness of a whisper or the percussive nature of a plosive, remained present in the absence of coherent language—as can be clearly heard in the audio examples below.

I treated the synthesis process as a bulk operation, using a variety of inputs for timbral transfer via the VAE-based RAVE models, and generating audio unconditionally via the latent diffusion models. While these latent diffusion versions produced output with higher clarity and more defined articulation, both functioned similarly within the project as sources of non-semantic speech where the resulting audio was treated as a singular musical palette.

Model Example 5.2 — *Smalltalk*: Model Audio^a

Examples of each trained model illustrate differentiated timbral characteristics: the whispered model ► shows a breathy, sibilant sound-quality; the ordinary speech model ► appears to preserve some harmonic structure and formants; and the overlapping (crowd) model ► seems to produce a broad, dense, speech-like texture.

^a. Portfolio/Chapter 5 - Smalltalk/Examples/raw model output/...

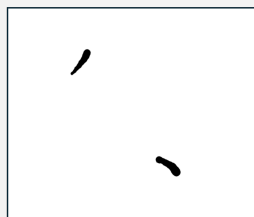
Synthesis to Organisation

In the composition, the selected segments of the generated audio were intended to be extremely short (often well below one second), and the success of this stage was evaluated by the tool's ability to parse the speech into usable micro-gestures. Representative audio examples in the box below illustrate this process: I would suggest that the retrieved segments broadly align with the intended contours, and show the use of the interface in the selection of compositionally useful material from a large audio corpus.

Portfolio Example 5.2 — *Smalltalk*: Contour Matching^a

Audio examples corresponding to a descending pair ► and longer descending figure ► show the method of contour-based selection. Further audio examples ► show manual selection from generated audio based on consonants.

Descending pair



Longer descent

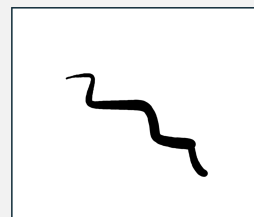


Figure 5.4: *Smalltalk* contour matching, in which simple shapes broadly correspond to retrieved audio in the *Portfolio*

^a. Portfolio/Chapter 5 - Smalltalk/Examples/similar shapes/...

Organisation to Device

As the motor movements which accompanied selected audio fragments were transcribed by hand, they consisted of a subjective comparison between the MIDI control data and the source audio. Figure 5.5 shows this process, where I transcribe the MIDI by ear from the arranged audio of the composition.

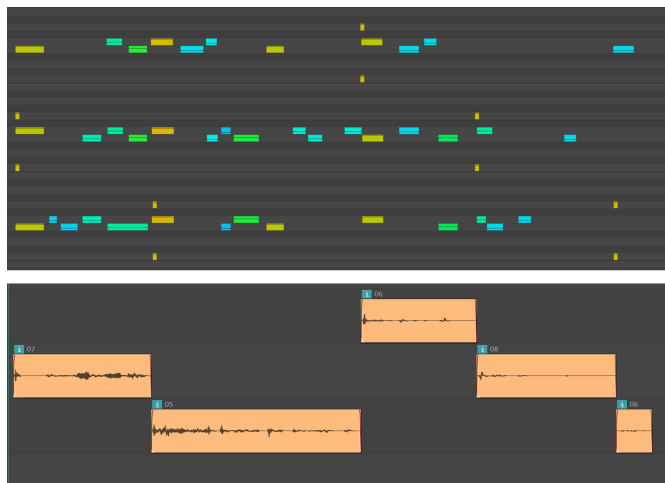


Figure 5.5: *Smalltalk* manually-transcribed MIDI (top) transcription from audio (bottom)

The transcription process highlighted several limitations in the project: with only two motors, and both using spring and metal plate combinations, the range of the mechanical aspect of the devices was particularly constrained. While the hand-transcriptions used a range of motor velocities, my view is that these were not particularly evident in the realisation. I see this as the approximate and ad-hoc qualities of the mechanism working against the motor's role in the composition—the speed of a motor might be controlled precisely, but the attached objects are not.

Despite this limitation, the objective of creating an apparent dialogue was achieved—primarily through the physical arrangement of the three devices. LEDs localised the sound, and while the mechanical constraints limited the expressive range of the work, the distribution of the three boxes established the basic conversational structure intended in the work's design.

Composition

The drawing interface allowed the creation of contour-related musical sequences that would have been time-consuming and impractical to find manually, and supported the form of the

composition as the presentation of disorganised and organised model outputs. However, my use of phonetic groups as a manual compositional strategy suggests that the tool might be more useful if extended with further selection parameters, incorporating timbre or texture rather than contour alone.

The piece functions as a demonstration of the contour-based compositional process, but the integration between the RAVE-generated audio and mechanical elements was dictated by physical constraints; the response of the spring-and-plate mechanisms lacked the nuance to fully shadow the micro-gestures retrieved by the DTW tool.

5.2.8 Summary

Smalltalk serves as a practical experiment in contour-based selection, with the DTW-based drawing interface a useful way to navigate an unlabelled, non-semantic generated corpus. While the asemic metaphor guided the synthesis and organisation stages, limitations in the physical devices revealed a threshold between the detail of digital audio and DIY mechanical response.

These findings inform the following work, *Radio User Guide*. While *Smalltalk* focused on the non-semantic contour of the voice through drawn shapes, *Radio User Guide* shifts the focus toward the gradual emergence of linguistic meaning, with a new approach to material organisation via a small text model, and an attempt to bind the material and computational aspects of the work more tightly through the associative qualities of the radio.

5.3 Radio User Guide

5.3.1 Overview

The second and final work in this chapter, *Radio User Guide* is an eight-minute piece for three players controlling disassembled radio and CD-players. To address R2 on the incorporation and organisation of RAVE¹⁶ generated audio, the audio in this work consists of the pre-recorded output of models trained on audio datasets of both speech and music—I use music here to link the speech audio of the work to its material aspects, through the theme of radio.

Over both Chapter 2 and Chapter 3 a small group of works were seen to deal with synthesised speech. Hecker’s *Inspection II*¹⁷ as one example presents a dialogue between synthesised voice and noise iteratively extracted from this source. Here, the ‘unexpected artefacts of synthesis’ are taken as musical material, and a similar interest is described by Walshe regarding *IS IT COOL TO TRY HARD NOW?*¹⁸ in the ‘weird, messed up sounds’¹⁹ of vocal synthesis²⁰. In both the generated speech and music of *Radio User Guide*, my focus throughout is on producing unstable or novel sounds, and highlighting these aspects within the piece.

In contrast to the aseptic approach of *Smalltalk*, the structure is rooted in the gradual emergence of semantic meaning. To achieve this, the form was produced by an LSTM text model trained on radio and CD-player manuals, which informed the organisation of musical material through intuitive word association. To highlight unstable and novel sounds, I trained audio and text models to different levels, shaping the work around a transparent presentation of these training processes. Model training processes have been explored in a range of computational art,²¹ and by creating a link between generated text and audio I explore how narrative can reframe abstract sound, attributing a causal relationship between the two.

16. Caillon and Esling, ‘RAVE: A variational autoencoder for fast and high-quality neural audio synthesis’.

17. Florian Hecker, *Inspection II*, 2019, accessed 25 February 2026, <https://lamp0.org/archive/florian-hecker-2019/>.

18. Jennifer Walshe, *IS IT COOL TO TRY HARD NOW?*, 2017, accessed 25 February 2026, <https://milker.org/jenniferwalsheworklist>.

19. Mark Dyer, Zubin Kanga and Jennifer Walshe, ‘In Conversation with Jennifer Walshe: Performing with Intelligent Machines’, *Contemporary Music Review* 42, no. 3 (2023): 391–399, <https://doi.org/10.1080/07494467.2023.2276563>.

20. Walshe’s work uses WaveNet—see Aaron Van Den Oord et al., ‘Wavenet: A generative model for raw audio’, *arXiv preprint arXiv:1609.03499* 12, no. 1 (2016), <https://doi.org/10.48550/arXiv.1609.03499>.

21. Ary-Yue Huang, ‘Artificial Life, One Leg at a Time: The Aesthetics of Trial-and-Error in AI Training’, in *Proceedings of the 2025 Conference on Creativity and Cognition* (2025), 818–821, <https://doi.org/10.1145/3698061.3726944>; Christian Sivertsen, Maarten Smith and Sander Van Der Zwan, ‘Art Critique by Other Means’, in *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (2023), 343–352, <https://doi.org/10.1145/3563657.3596069>.

Objectives

1. **Dataset:** Curate audio of audiobooks, online dictionaries, and music, to provide a range of material to train RAVE models.
2. **Synthesis:** Train several RAVE models on each dataset, to different levels.
3. **Device:** Construct three devices from found objects, collected from discarded radio and CD-players. Control each via a MIDI keyboard.
4. **Organisation:** Train a text model on radio and CD-player manuals, to provide the form for the work.
5. **Composition:** Follow the output of the text model, using word association to arrange the generated audio, linking the text, speech and music together.

5.3.2 Dataset

I trained one speech and five music RAVE models for this project. For the first speech model, a dataset of audio was created by scraping online dictionaries and archives for words occurring in a radio manual vocabulary. This was to allow for the presentation of gradual training on specific words. A second speech model was reused from the previous project, *Smalltalk*, trained on audiobooks.

Model Example 5.3 — *Radio User Guide: Dataset*

The dataset of radio-manual words^a ☐ scraped from online dictionaries is made available. Each word was overlaid with musical extracts from the open-source FMA-large dataset, to simulate radio clips.

^a. Models/Chapter 5 - Radio User Guide models/RAVE models/allWordsAudio.wav

Music models were trained on genres from the open-source FMA-large dataset.²² My intention of the genre classification was to ensure variety through different model outputs. However, the range of material in each category remained wide enough to produce unexpected audio—for example, the ‘classical’ model dataset contained piano, choir, and orchestral music.

22. Michaël Defferrard et al., ‘FMA: A dataset for music analysis’, *arXiv preprint arXiv:1612.01840*, 2016, <https://doi.org/10.48550/arXiv.1612.01840>.

5.3.3 Training and Synthesis

RAVE models are trained in a two-step process:²³ first, the model acts as a standard Variational Autoencoder (VAE) with the goal to compress audio into a compact latent space and reconstruct it accurately. The second stage uses a Generative Adversarial Network (GAN) framework to improve the synthesised audio quality.

In this creative project, speech and music RAVE models were trained to varying degrees of completion, saved after one hour of training, then two, six, and so on. As with *Smalltalk*, I chose the training configuration through trial-and-error adjustment. All models were trained with ‘v2’ and ‘wasserstein’ configurations, without augmentation. RAVE Latent Diffusion²⁴ models were also trained in increments, with examples exported between training.

The result of this process was that earlier models often have a ‘noisier’ quality, and the purpose of generating audio with these earlier versions was to emphasise the unstable and noisy aspects of the output. The process of gradual convergence, from noise to recognisable sound, becomes the overall form of the work.

Model Example 5.4 — *Radio User Guide: Models*

Each model is made available.^a □ There are several versions of each model, from poorly to better-trained.

^a. Models/Chapter 5 - Radio User Guide models/RAVE models/...

5.3.4 Device

In *Radio User Guide*, the radio is only relevant through an emphasis on the physical and associative qualities of the materials; the speech and music audio generated for this project have no origins in radio broadcasts, but when combined with the discarded parts of a radio and a nonsense instructional text, a connection is made regardless.

To create the physical devices, I took parts from three broken radio and CD-players, using the motors, wires, sections of plastic casing, PCB, and gears, and speakers. Each device used a raspberry pi, contact microphone, and MIDI keyboard. For audio sample playback, a single speaker was connected to the Pi, with the other connected to a contact microphone on the surface

23. Caillon and Esling, ‘RAVE: A variational autoencoder for fast and high-quality neural audio synthesis’.

24. Horta Valenzuela, *Rave Latent Diffusion*.

(Figure 5.7). Components were glued to clear Perspex, with holes cut for each speaker—I see this as a continuation of transparency from the computational to the material (Figure 5.6).

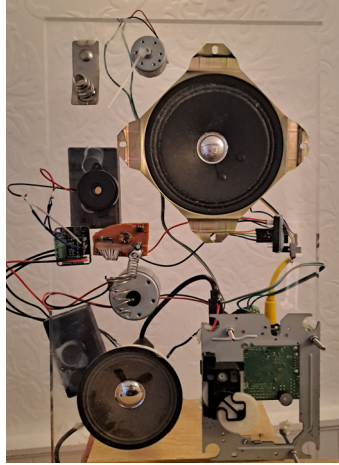


Figure 5.6: One of three devices

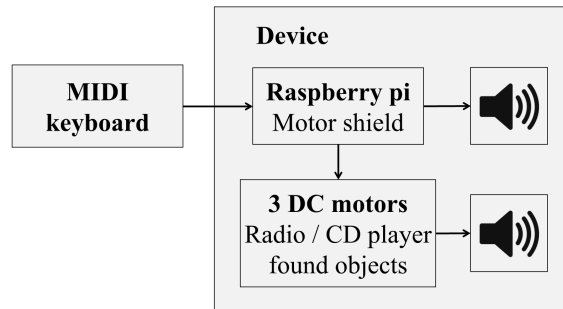


Figure 5.7: *Radio User Guide* design

The sound of each device was the result of experimentation with a range of objects from the CD and radio players, from the scraping and clicking of battery compartment springs, plastic and cable ties, to the sounds of the gears and motors. To produce a broadly uniform noise level when amplified, motor speeds were mapped from MIDI velocity, adjusted by ear.

5.3.5 Organisation

To connect the trained models and devices, the method I used to organise generated material draws on existing works that use algorithmic methods of textual distortion and fragmentation as a compositional strategy, from Harald Muenz’s *deChiffrAGE*²⁵ to Timothy Polashek’s *Beyond Babble*.²⁶ More recently, works like Kirath’s *Cacaphonic Choir*²⁷ and Jennifer Walshe’s *Text Score Dataset*²⁸ have used language models to expand this practice. The advantage of using a text model is that where results of an algorithmic shuffling might be more predictable, I found that deliberate under-training of text models can produce unexpected results.

25. Harald Muenz, *deChiffrAGE*, 2004, accessed 25 February 2026, <https://www.musicwedliketohear.com/pdf/29032006.pdf>.

26. Timothy D Polashek, ‘Beyond Babble: A Text-Generation Method and Computer Program for Composing Text, Music and Poetry’, *Leonardo Music Journal* 15 (2005): 17–22, <https://doi.org/10.1162/lmj.2005.15.1.17>.

27. Şöl en Kirath, Hannah E Wolfe and Alex Bundy, ‘Cacophonic choir: an interactive art installation embodying the voices of sexual assault survivors’, in *ACM SIGGRAPH 2020 Art Gallery* (Association for Computing Machinery, 2020), 446–450, <https://doi.org/10.1145/3386567>.

28. Jennifer Walshe, *The Text Score Dataset 1.0*, 2021, accessed 25 February 2026, <http://milker.org/s/The-Text-Score-Dataset-10-Jennifer-Walshe.pdf>.

To provide the form for the work, I trained a character based LSTM model on several radio manuals, augmenting the dataset through sentence shuffling to offset the small size. My decision to use an LSTM network rather than a transformer in this work was prompted by the small vocabulary of the dataset making a more powerful model unnecessary. While lightweight LSTM networks can struggle with long term dependencies, that was not a concern in this work as errors in early training were part of the structure.

Every five epochs during training, a version of the LSTM was saved. Fourteen were chosen for the final work, showing a gradual progression towards coherent text (Table 5.2).

Table 5.2: Example LSTM text outputs.

Epoch	Text Example
1	.wUb, nh SeaDsNrr lrcDm:loABc5mCzTxHd3k2ks
6	e e n ee t e e e e o en ee
11	r t lo aao r t i n ai ae iso r r actao t on t oe t
26	nd to the to the tor the the to the on the the to the the
61	diss the dered the mord the button the button the unit
81	and the maissit the prockect the unit the to the unit the prory the cassette the FUNT button
101	Press the cassette to the promect the dispose the cassette into the press the mains and the unit on the
201	to station the station to playback will station of the disc and the disc track and the disc on the CD door from
401	To selected to record in the unit off and place the unit is not open the program button to store discs or the tabs
1001	using the unit on to a third party make sure the unit is now connecting the unit is now connecting the unit can be
2001	2. Fit two batteries into the lower battery compartment, observing polarity as shown. 4. Re fit the battery life.
2501	1. Unclip the two battery cover latches and remove the battery cover 2. Fit two batteries into the lower battery leakage and corrosion or battery...
3001	Failure to do so will risk severely damaging the disc off the turntable and place it back in the case with a slightly moist, lint free cloth. Do not use any cleaning
4996	Press the program button to store the track number. 4. Repeat steps 3 to 4 to program other desired tracks. 5. Press the CD PAUSE button, playback will start from the

Particular words suggested actions:²⁹ the appearance of the word ‘stop’ or ‘pause’ caused players to rest; ‘repeat’ or ‘cycle’ caused a loop of previous material. This was to produce a meaningful association between the spoken words in the piece, and the surrounding audio or motor activity (Table 5.3). This list of word associations produced the initial form of the work, but remained flexible during the editing process.

Table 5.3: *Radio User Guide* word association examples.

Loop	Pause	Radio	Motors
<i>memory</i>	<i>pause</i>	<i>play</i>	<i>unit</i>
<i>cycle</i>	<i>stop</i>	<i>track</i>	<i>equipment</i>
<i>previous</i>	<i>off</i>	<i>tracks</i>	<i>device</i>
<i>record</i>	<i>cancel</i>	<i>tune</i>	<i>rotating</i>
<i>repeat</i>	<i>wait</i>	<i>radio</i>	<i>apparatus</i>
<i>again</i>	<i>eject</i>	<i>playback</i>	<i>enclosure</i>
<i>store</i>	<i>interrupt</i>	<i>tape</i>	<i>compartment</i>
<i>...</i>	<i>disconnect</i>	<i>cd</i>	<i>tray</i>
	<i>...</i>	<i>...</i>	<i>...</i>

Portfolio Example 5.3 — *Radio User Guide*: Word Association^a

While many examples were drawn from the initial table, I decided others intuitively.

- **the (1-5)** show the gradual emergence of the word ‘the’ through different model training stages, in the context of the composition. 1. ▶ 2. ▶ 3. ▶ 4. ▶ 5. ▶
- **skip (1-3)** I decided would be associated with a rapid sequence of new material. These examples also show the progression through better-trained music models. 1. ▶ 2. ▶ 3. ▶
- **recall** was associated with the recurrence of a previous sound. ▶
- **swallowed** occurs once, where I use a speech model to transfer the timbre of an orchestral extract to voice. ▶

^a. Portfolio/Chapter 5 - Radio User Guide/Examples/...

29. My approach to linking narrative and sound here was informed by Tom Johnson’s *Bedtime Stories* (1985) in which a narrator completely reframes clarinet melodies—showing how playful and surprising connections between narrative and abstract sound can be.

5.3.6 Composition

The compositional process involved layering the mechanical sounds of the devices with the generated audio clips, following the structure provided by the text model.

Collage

I created the opening section of the work by collaging short clips of generated audio from poorly-trained speech and music models. My approach was to present a large amount of material in rapid sequences, drawing a broad analogue to the model training process. Inputs to the models included noise, and fragments of the original datasets.

Speech in the first section used the word *the* as input. The use of this word, alongside generated noise from the music models, corresponded to an early training stage of the text model, which produced a repetitive stream of *the* separated by nonsense words:

nd to the diss the tor the the mord the on the the dered the the...

I found the generated text of the early LSTM versions pleasingly similar, on a surface level, to early examples of sound poetry.³⁰ Beginning with this material was a way to emphasise a gradual progression—to better-trained models and longer audio fragments aimed at creating an emergence of intelligible speech.

Portfolio Example 5.4 — *Radio User Guide: Collage*^a

- **collage 1** ► and **collage 2** ► highlight the general approach, using several short clips in fast sequences.
- **collage 3** ► shows, to me, that even a thirty-millisecond second clip of poorly generated speech is recognisable as a human voice.

^a. Portfolio/Chapter 5 - Radio User Guide/Examples/...

Feedback Loops

Sending the output of a model through itself, or another model, was a way to emphasise the particular qualities of a model and produce degraded and noisy material. In my thinking, this

30. Raoul Hausmann, *f m s b w t ö z ä u*, 1918, accessed 25 February 2026, <https://www.centrepompidou.fr/fr/ressources/oeuvre/cozKBB>.

was also linked to the word associations—for example, the word ‘corrode’ triggered a process of running the audio through progressively earlier versions of the same model, resulting in a subtle degradation of the sound.

Portfolio Example 5.5 — *Radio User Guide: Feedback Loops*^a

- **isRecording** feeds the output of the model back into the model. This highlights some features of the ‘electronic’ model, like the tendency to turn any percussive sound into an electronic pad. ►
- **again** sends the output of a model into a different model. The differences in the qualities of each model can be heard more clearly in this case. ►

^a. Portfolio/Chapter 5 - Radio User Guide/Examples/...

Scoring

In performance, each device was controlled using a MIDI keyboard (Figure 5.8). One note triggered the next sound sample in a list, and three notes controlled the motors directly. To rehearse the piece, a set of notes were used for rehearsal letters, and a single note was used to reset the audio samples to the start. Volume was also adjusted using a key, to balance the audio with the motors.

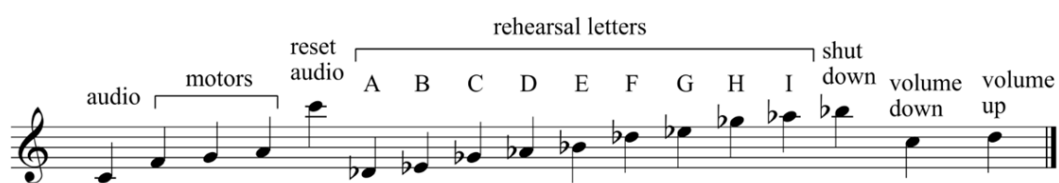


Figure 5.8: *Radio User Guide* score key

Portfolio Example 5.6 — *Radio User Guide: Performance*

Aspects of the audio recording are highlighted below,^a ► with the score presented in the portfolio.^b ≡ For a one-minute extract showing the visual aspects of the work, see the video example.^c ►. Code for performing the work is also made available. </>

- 0:00 - 01:20 plays noise through the different models, in collaged sequences.

- 01:20 - 02:20 uses the word ‘the’ with poorly trained RAVE models (following the output of the text model), with 01:32 an example of noise played through an early speech model.
- 02:20 - 03:50 shows the gradual emergence of intelligible speech. While the text is nonsense, recognisable words and phrases begin to appear.
- 03:50 - 05:45 uses clearer examples of speech, and more complete phrases from the later text models.
- 05:45 presents clear speech and phrases from the text and audio models, with 06:53 an example of piano through the classical model. In this final section, my focus is on associations between the text and audio.

a. Portfolio/Chapter 5 - Radio User Guide/Radio_User_Guide_audio.wav
b. Portfolio/Chapter 5 - Radio User Guide/Radio User Guide - score.pdf
c. Portfolio/Chapter 5 - Radio User Guide/Video_example.mov

5.3.7 Evaluation

In this evaluation I examine the incorporation of RAVE and LSTM models into the electromechanical practice, focusing on how the organisational methods managed the structuring of generated material, and how generated audio was integrated with the physical devices.

Dataset to Synthesis

The specific characteristics of each dataset influenced the character of the RAVE outputs. For example, the dictionary-scraped dataset in particular—with its overlaid fragments of audio—produced synthesis that initially improved with training, then remained in a relatively noisy state regardless of training duration (Portfolio Example 5.7). This provided a consistent source of fragmented speech-like audio that lacked the polished qualities of the reused *Smalltalk* model.

Similarly, the diversity of the FMA-large music subsets resulted in unexpected timbral qualities, which I interpret as the models attempting to reconcile disparate features from the dataset, such as the combinations of piano, choir and orchestral colour found in the *Classical* model.

However, one limitation was the trial-and-error nature of model training: I could not predict the output until several days of training had passed. This meant training several models in parallel, which was a long and inefficient process.

Portfolio Example 5.7 — *Radio User Guide: Dataset to Synthesis*

- Dictionary-speech model: Speech audio not in the original dataset (audio start) followed by the model's rendering at 1, 12, 24 and 36 hours of training demonstrates initial success in reconstruction, but a lack of progress and retention of artifacts.^a ►
- Timbral qualities: The diversity of each dataset led to several music models producing audio with unexpected timbral combinations, such as the *Classical* model oscillating between choir and piano.^b ►

^a. Portfolio/Chapter 5 - Radio User Guide/Examples/speech training.wav

^b. Portfolio/Chapter 5 - Radio User Guide/Examples/ChoirAndPiano.wav

Synthesis to Organisation

My aim with the work was to make the progression of model training visible through the structure. A limiting factor in this respect was the recognisability of poorly trained generated speech, as speech. To correspond with the semantic fragmentation defined by the LSTM, I used other techniques—either millisecond clips, using the music models to reproduce speech, or obscuring the speech with other sounds. This allowed the uncanny effect of the fragmented text to persist longer than the RAVE synthesis allowed on its own.

In my view, the under-trained LSTM text model provided a structure of fragmentation that proved more interesting than rule-based methods. By using this method, the different versions of the text model produced a natural emergence of coherent text in a way that mirrored the trajectory of the audio models, as speech and music gradually emerge from noise.

Model Example 5.5 — *Radio User Guide: Synthesis to Organisation*

Each version of the text model is made available with a script for testing.^{a,b} With the audio models, speech was instantly recognisable even when poorly trained, as in this example after one hour of training.^c ►

^a. Code/Chapter 5 - Radio User Guide code/radio model DEMO.py

^b. Models/Chapter 5 - Radio User Guide models/Text models/...

^c. Portfolio/Chapter 5 - Radio User Guide/Examples/obviousSpeech.wav

Organisation to Device

The role of the motor-driven elements was less prominent than works in the previous and subsequent chapters, perhaps a consequence of the wide range of synthesised material produced for the work. However, by writing the piece in such a way that connected the text-model output to motor actions I intended to create a perceived causal link between the machine-generated narrative and the physical hardware, with the motors functioning either as punctuation for the digital audio or responding to associated words in the text (such as **APPARATUS** or **UNIT**).

Aside from these associative qualities, the physical distribution of the work across three devices allowed for a spatial fragmentation of the material, adding to the already fragmented output of the models and their compositional arrangement. Connections between the models and physical hardware were extended with the variability of the physical objects suggesting the instability of the models; with each MIDI trigger, the devices produced inconsistent sounds, connecting the electronic audio to the behaviour of unstable mechanical systems.

Composition

While the LSTM provided a form, the resulting piece and its transition from noise to intelligibility was composed around this, not automatically generated. My reliance on manual assembly, collage and feedback loops highlights the limits of the organisational method. For example, feedback loops were used to deliberately degrade the RAVE output—and the final musical narrative relied on the curation of model failures rather than simply the raw output of the training process.

This approach could be developed in several directions, with possibilities in live generation by using the text model to direct the work rather than as a pre-compositional step. While a real-time algorithmically-driven form is appealing to me, the use of human players and pre-composed material allowed a detailed exploration of generated material, and compositional intervention and play with the output of the text and audio models was an important part of the creative process. There are however possibilities for involving human feedback or disruption into the system, to connect the roles of humans and machines in performance more fully.

5.3.8 Summary

In *Radio User Guide* I have explored the incorporation of speech-trained audio and text models into an electromechanical practice involving found objects and human performers. By using iterative model training, the work presented the gradual emergence of coherent speech organised around machine-generated text.

One particular challenge I found in working with generated speech was managing its high recognisability, which required manual intervention to maintain a level of fragmentation. In my view, this balance between automation and intuitive composition was required to ensure the form remained transparent. The process resulted in a piece which moved from digital and mechanical noise towards linguistic coherence, where the motor-driven parts functioned both as punctuation, and in response to the synthesised audio.

5.4 Discussion

This discussion addresses R2, on the integration of NAS generated audio into the electromechanical practice, and the methods developed to organise this material. This section is structured in three parts, relating to the objectives of the chapter: Generating Instability, Methods of Organisation, and Material Integration.

5.4.1 Generating Instability

Recorded speech is usually bound to the identity of a speaker.³¹ By using a large dataset and approximate transcription in the previous chapter, I began to disconnect the individual speaker from the musical output, to turn the focus to rhythmic, amplitude or textural aspects—features of speech to be developed musically. Likewise, working with RAVE models detaches prosody, timbre, and phonetic components from a human source, allowing them to be treated as sonic material apart from contextual cues. Its affordances in my creative practice are the malleability of the output, as the material can be generated in bulk to fit specific parameters with a connection to the timbral qualities of speech, apart from meaning.

In my view, this detachment is central to the technological uncanny; the models produce audio that is recognisably speech-like, but is ultimately a decontextualised and flattened repres-

31. Denis Smalley, 'Defining Transformations', *Interface* 22, no. 4 (1993): 279–300, <https://doi.org/10.1080/09298219308570638>.

entation of the source. I used specific methods to amplify these qualities in each work. In *Radio User Guide*, feedback loops were used to emphasise the characteristics of each model—such as forcing the electronic model to converge on percussive, drum-pad-like sounds—while timbral transfer allowed for the use of non-speech inputs to impose vocal qualities onto non-human sources in *Smalltalk*.

Transcribing the resulting audio to motors extended this process by emphasising features like amplitude or rhythm while discarding others. In *Smalltalk*, this took the form of a gestural shadowing of the audio, while in *Radio User Guide*, I paired the artifacts of the under-trained models with the unstable mechanical noise of found radio parts.

5.4.2 Methods of Organisation

The methods of organisation in this chapter were a response to the volume of audio produced by neural synthesis. While the transcription process in Chapter 4 was about mapping a fixed source, the generative nature of RAVE involved a shift toward selection strategies. In both *Smalltalk* and *Radio User Guide* I used automated and manual methods to organise generated audio, targeting different aspects of the material and establishing the basis for each composition (Table 5.4).

Table 5.4: Organisational Methods in *Smalltalk* and *Radio User Guide*

	Smalltalk	Radio User Guide
Selection Basis	Asemic, audio pitch and amplitude contour.	Semantic, model progression from noise to coherence.
Selection Unit	Individual micro-gestures	Training stages (epochs) representing model progress
Automation Role	Matching drawn input to specific audio fragments.	Generating text to define musical structure.
Manual Role	Manual selection of returned segments.	Manual assembly and collage around generated form.
Resulting Structure	Study of repeating shapes and patterns.	Linear progression of emerging semantic meaning.

In *Smalltalk*, the drawing interface functioned as a way to segment and group audio by pitch and amplitude contours. This could be seen as allowing the work to be constructed from the bottom up, where the primary organisational task was the identification and grouping of individual micro-gestures. In *Radio User Guide*, the organisation moved away from retrieval and

toward form, by using an LSTM text model to generate a script for the work. In this instance, the organisation was a text-based scaffold for the placement of audio and mechanical actions.

In both works, the organisational method provided me with a specific compositional constraint. In *Smalltalk*, the constraint was on the features of the musical material, focusing the composition on the repetition and variation of contour. In *Radio User Guide*, the constraint was narrative, focusing the composition on the emergence of intelligible text and speech over time. These methods defined the specific parameters through which the generated audio could be meaningfully composed.

5.4.3 Material Integration

In contrast to the timbrally rich and varied generated audio, the electromechanical actuators had an extremely limited range. This mismatch between the digital and the physical shaped the compositional process in both works, but was taken to an extreme in *Smalltalk*, in which I found the difference between the audio and mechanical elements of the work were never properly reconciled.

In *Radio User Guide*, I addressed this by limiting the activity of the motors to specific sections—in an accompanying role—and integrating the motors conceptually through the material aspects of the work. I would suggest that this integration was more effective because it moved away from direct mimicry, and by writing the motor parts in response to keywords from the text model they were part of a broader narrative.

The integration in *Radio User Guide* was further supported by the diversity of the audio material. *Smalltalk* focused strictly on generating speech, which placed the motors in a direct and unachievable comparison with the voice. *Radio User Guide* instead incorporated music and noise, and the broader palette allowed the mechanical rattling of found radio parts and springs to sit within a dense field of sound where they were not the sole focus of representation. The found objects also carried contextual cues that aligned with the broader theme of the radio; the mechanical noise became an extension of the garbled instructional text rather than an imitation of the human voice.

Both works in this chapter are characterised by this combination of cutting-edge NAS and lo-fi mechanical aspects, but in *Radio User Guide* these contextual cues suggest correspondences between old and new technology—from the noisy aspects of the audio produced by the poorly-

trained models and the noise of radio static, to the processes of information compression and transmission that are central to each.

5.4.4 Conclusion

In this chapter I have detailed the process of incorporating speech-trained RAVE models into an electromechanical practice through training, generation, selection, and physical realisation.

The primary finding relates to the organisational methods applied to the models' output through two distinct compositional strategies. The first, in *Smalltalk*, used contour-based searching to discover musical material based on visual-sonic similarity. The second, in *Radio User Guide*, involved deriving the structure of the work from generated text.

For me, the humour and absurdity that results from the process of fragmenting and recombining language was a large part of each work. In both cases, by detaching prosody and timbre from human identity, the models produced an uncanny—and often comic—form of fragmented speech that functioned as a raw material for the electronic and motor-driven compositional practice. Reflecting on the integration of the electromechanical elements, I would suggest that the substantial loss of information during the translation from speech features to motor control resulted in a noticeable disconnect in *Smalltalk*, while in *Radio User Guide* these elements were more tightly bound through the material and conceptual logic of the work.

The following chapter returns to motors as central to the realisation of speech-derived musical material, by focusing on syntax patterns as a driver of musical structure.

Chapter 6

Text Models in Practice

6.1 Introduction

In this chapter I present three works using text models as the basis for generating musical material. This work represents a further abstraction from the recorded and generated speech of Chapters Four and Five, to focus specifically on the use of text models as generators of linguistic structure: In *Porcelain Music Box*, I take an LSTM trained on dialogue acts as the starting point for a small-scale installation work, and in *Seven Studies* I use a small language model to produce dependency relations, forming the basis of a modular mechanical improvisation system. In the final work, *Word games*, I use word embeddings to generate musical material.

Chapter 2 found linguistic features to be an underexplored area of speech-derived practice. While examples exist—Bosetti’s *Mask Mirror*¹ uses linguistic categories as the basis for triggering random sequences of speech audio—these often focus on the meaning of language. The abstracted use of linguistic structures is less common, with Barlow’s *Prog  thal Percussion*² as a notable example, mapping part-of-speech categories to a musical score for six percussionists with a fixed set of rules. More recently, Ko et al’s *Harmonic Words, Narrative Chords*³ presents a method for translating part-of-speech categories into chord sequences.

1. Alessandro Bosetti, *Mask Mirror*, 2008, accessed 25 February 2026, <https://melgun.net/travaux/maskmirror/>.

2. Clarence Barlow, *Prog  thal Percussion for the Advanced Beginners*, 2003, accessed 25 February 2026, <https://clarlow.org/compositions-by-year/>.

3. Jiin Ko, Youngjun Choi and Jinjoon Lee, ‘Harmonic Words, Narrative Chords: Textual Sonification Using Part of Speech’, in *Proceedings of the International Conference on New Interfaces for Musical Expression* (2024), 28–32, <https://doi.org/10.5281/zenodo.13904771>.

I use text models here to extend existing methods, such as Barlow’s ‘semantic composition’.⁴ Rather than taking an existing text, this allows linguistic structures to be generated and mapped in real-time. The final work, *Word games*, is an exception, in which I map an original, fixed text. Throughout Chapter 4, the focus of the mapping of speech audio to motors was on the abstracted qualities of speech, rendered through found objects. Here I apply the same principle to linguistic speech structures, where the organisation of language forms the basis for each work, with the aim of addressing R3:

R3. What are the affordances and limitations of text models as creative tools, applied to linguistic speech structures?

Objectives

1. **Devices:** Build three DIY devices for text-driven sound production.
2. **Linguistic Mapping:** Translate dialogue acts, dependency relations, and word vectors into musical frameworks.
3. **Interaction and Control:** Design interfaces for automated installation playback and real-time improvisation.

These objectives are realised through the three works outlined in Table 6.1. While in *Porcelain Music Box* and *Seven Studies* I use generative models, in *Word games* I apply similar mapping strategies to static word embeddings.

Table 6.1: Overview of Project Elements: Text Models in Practice

	Porcelain Music Box	Seven Studies	Word games
Linguistic Structure	Dialogue acts	Dependency relations	Word embeddings
Text Model	LSTM	Small Language Model	—
Device	12 DC motors, porcelain fragments	16 DC motors, modular found objects	9 DC motors, modular found objects
Creative Output	Installation (automated)	Improvisation (real-time)	Pre-composed

4. Clarence Barlow, ‘On Music Derived from Language’, in *Proceedings of the 3rd International Symposium on Systems Research in the Arts and Humanities, Baden-Baden, Germany* (2009), 4, accessed 25 February 2026, <https://clarlow.org/wp-content/uploads/2025/06/On-Music-Derived-from-Language.pdf>.

6.2 Porcelain Music Box

6.2.1 Overview

Porcelain Music Box is a small-scale installation piece that explores conversational structure as the basis for generative musical composition. The work features twelve DC-motors mounted on a wooden box, each striking a fragment of broken porcelain.

In response to Chapter 2 finding grammar structures to be an underexplored source of speech-derived music, I began this work with a dataset of dialogue acts (see 6.2.2). At the core of the piece is a generative system, driven by a text model trained on dialogue-act-labelled transcripts of human conversation. Rather than producing text or speech, the model generates sequences of dialogue acts such as *statement*, *question* or *backchannel*, and translates them into motor actions, determining which motors are triggered, in what order, timing, and dynamics.

Objectives

1. **Linguistic Structure:** Generate sequences of dialogue acts.
2. **Mapping:** Translate dialogue acts into musical patterns based on concepts of continuation and disruption.
3. **Device:** Build a DIY device of twelve DC motors and porcelain fragments to produce a constrained acoustic palette.
4. **Interface:** Design a minimal, audience-triggered controller for automated playback.

6.2.2 Linguistic Structure

Dialogue Acts

Dialogue acts are the functional units of conversation that represent a speaker's communicative intention,⁵ such as making a statement, asking a question, or a backchannel—utterances like *uh-huh*—which continue the flow of a conversation during another speaker's turn. These labels are used to understand the functional intent of speakers, segmenting dialogue into discrete events

5. Harry Bunt, 'Dimensions in Dialogue Act Annotation.', in *Proceedings of the Fifth International Conference on Language Resources and Evaluation* (European Language Resources Association, 2006), 919–924, accessed 25 February 2026, http://www.lrec-conf.org/proceedings/lrec2006/pdf/428_pdf.pdf.

based on their interactional role.⁶

In my view, this categorisation makes dialogue acts valuable material for an algorithmic musical system. Rather than attempting to process or respond to the full complexity of the annotation, in the system developed for *Porcelain Music Box* I use the structure and sequence of dialogue acts as a source of control data. Their categorical nature allows them to be mapped directly to discrete motor sequences, simulating patterns in a conversation.

To implement this approach, I chose the Meeting Recorder Dialog Act corpus⁷ (MRDA) as the training dataset. This consists of multi-party conversations of real-world meetings, with each utterance annotated for speaker identity, dialogue act, and subcategories.

For this project, I chose to use the speaker identity and main dialogue act category to keep the work focused on higher-level patterns rather than the myriad subcategories of dialogue act.⁸ To preprocess the data, conversations with more than six participants were filtered out, and the remaining sequences were used to train a model that predicts the likely next speaker and corresponding dialogue act. These predictions formed the basis of the system, allowing for a conversationally-informed musical output.

Model

I trained an LSTM model to generate sequences of dialogue acts, with a total of six speakers (speaker ID's zero to five). This approach was adapted from a standard architecture used for lightweight text generation:⁹ the embedding layer converted input tokens into vector representations, followed by two LSTM layers. A final dense layer produced predictions over the vocabulary. I converted this to a tf.lite model,¹⁰ to run on the device in real time.

An example of the output might be a sequence of statements from speaker zero, with backchannels from speaker three:

6. Andreas Stolcke et al., 'Dialogue act modeling for automatic tagging and recognition of conversational speech', *Computational linguistics* 26, no. 3 (2000): 339–374.

7. Elizabeth Shriberg et al., 'The ICSI meeting recorder dialog act (MRDA) corpus', in *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004* (2004), 97–100, <https://doi.org/10.21236/ada460980>.

8. Shriberg et al.

9. Daniel Jurafsky and James H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*, 3rd (2026), 13, accessed 25 February 2026, <https://web.stanford.edu/~jurafsky/slp3/>.

10. Martín Abadi et al., 'TensorFlow: A system for Large-Scale machine learning', in *12th USENIX symposium on operating systems design and implementation (OSDI 16)* (2016), 265–283, <https://doi.org/10.48550/arXiv.1605.08695>.

0S 0S 3B 0S 0S 3B 0S 3B 0S 0S

This appears to be a particularly sparse data source. To translate these sequences into musical patterns, I decided each speaker ID would determine a pair of motors, which moved according to predefined rules (see *Mapping* below). The structure of the program can be seen in Figure 6.1.

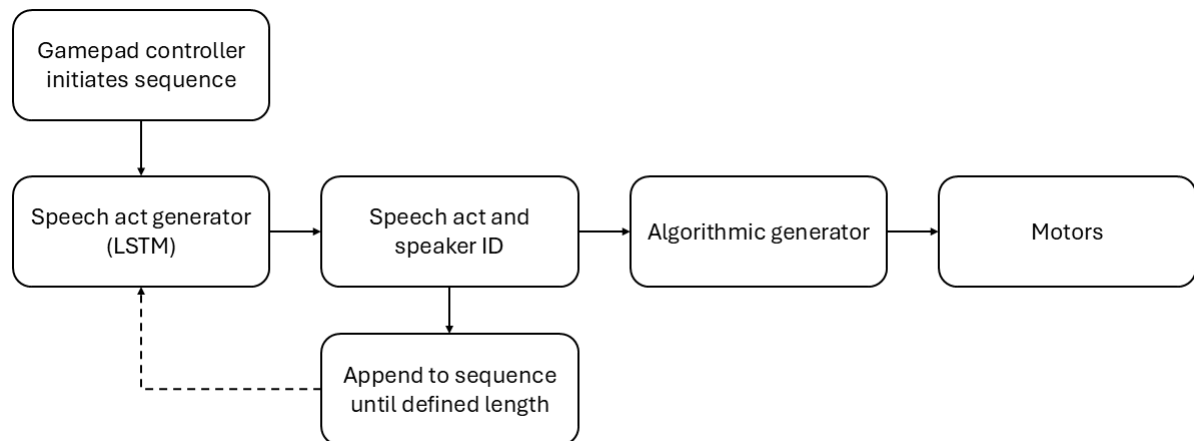


Figure 6.1: *Porcelain Music Box* program structure

Model Example 6.1 — *Porcelain Music Box*: Model

The LSTM, tokenizer and vocabulary are made available.^a 

^a. Models/Chapter 6 - Porcelain Music Box models/...

6.2.3 Mapping

While each of the five dialogue acts have a particular conversational function, the categories could also be seen as metaphorically describing continuations or disruptions to a conversational flow: a backchannel continues a conversation, while a disruption interrupts.

This led to the use of a fluctuating wave in the mapping process, with corresponding continuations and disruptions. Groups of statements and backchannels were seen as continuations: a sequence of gradually increasing or decreasing durations and amplitudes were generated along a log scale specifying the start, end, and number of intermediate points (Figure 6.2, 1a). A disruption or floorgrabber however would interrupt the sequence at a new point (Figure 6.2, 1b).

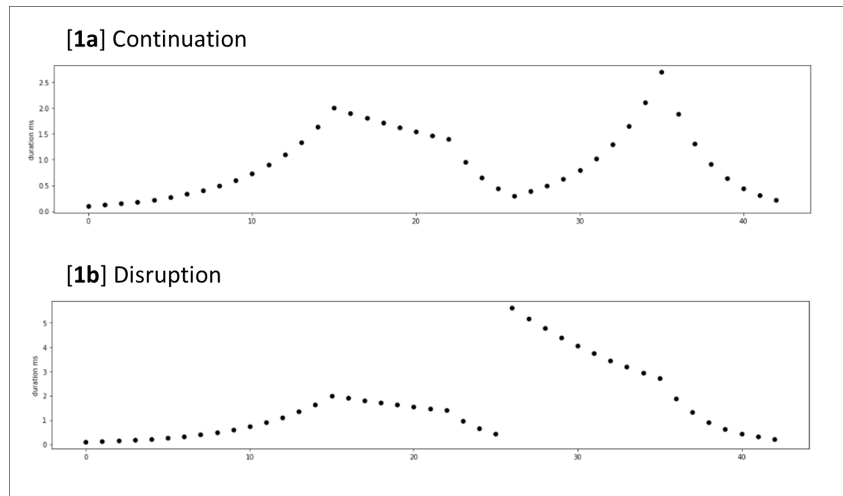


Figure 6.2: Patterns of Continuation and Disruption

Some acts had extra qualities, which were intuitively decided but informed by the conversational source: a floorgrabber always started short and loud, gradually becoming shorter and softer; questions always gradually became faster, followed by a slowing down; if a statement followed a question, it would also use the same motor order, but with a new rhythm and amplitude. To slightly disrupt these patterns, grace notes were added, and the overall speed of events modulated gradually.

The detail of the work, the start and end points, grace notes, speed modulation and so on were generated using a Python script, which set new durations and amplitudes as required by the output of the model.

Code Example 6.1 — *Porcelain Music Box*: Script

The iteration of the Python script shared here was used for testing rather than the final configuration of Arduino and motor boards, and therefore may be of more use. This sends the output of the model and mapping via OSC to Max-MSP.^a </>

^a. Code/Chapter 6 - Porcelain Music Box code/Porcelain Music Box code.py

6.2.4 Device

The function of the device was to generate motor sequences of percussive sound, through the porcelain fragments on the surface. My design of the device was informed by its setting: this is the only work in this project without amplification, as the work was for a small installation in

a library, and the porcelain did not require amplification to be audible.

I chose CD-drive motors as they were relatively quiet, and low cost. A Raspberry Pi was chosen to run the text model using a Python script, and send serial data to three Arduinos which in turn each controlled four motors (Figure 6.3). All components were placed into a small wooden box.

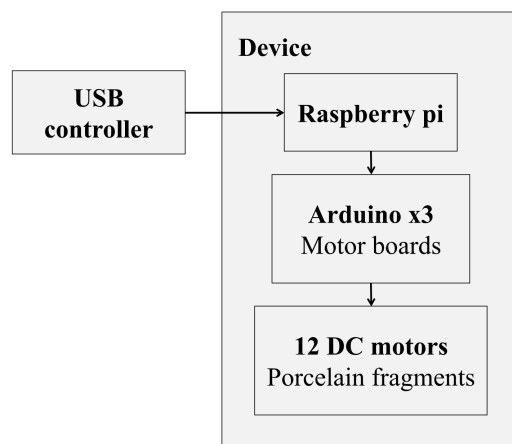


Figure 6.3: *Porcelain Music Box* design

The porcelain pieces were taken from a potters studio, prior to the work. I chose specific fragments for their pitched qualities, and fixed these to the box surface. Two guitar strings were also stretched over a larger piece of porcelain, which produced a new sound quality, and increased the resonance of the device. Each motor spun a cable tie, which struck either the porcelain or the guitar strings (Figure 6.4).

6.2.5 Interface

Unlike the *Motorcase* or other improvisation-driven devices, this was a small installation work for interaction by the public, and so I developed a minimal interface around a USB gamepad controller. While this interface was originally intended to use only two buttons, I felt that the additional buttons of the gamepad demanded some kind of associated action. These additional buttons responded directly to the controls, by sweeping through the motors in certain directions.

The controls for this project were deliberately minimal to suit the installation setting:

- **Four directional controls:** sweep through the motors in the direction indicated with a

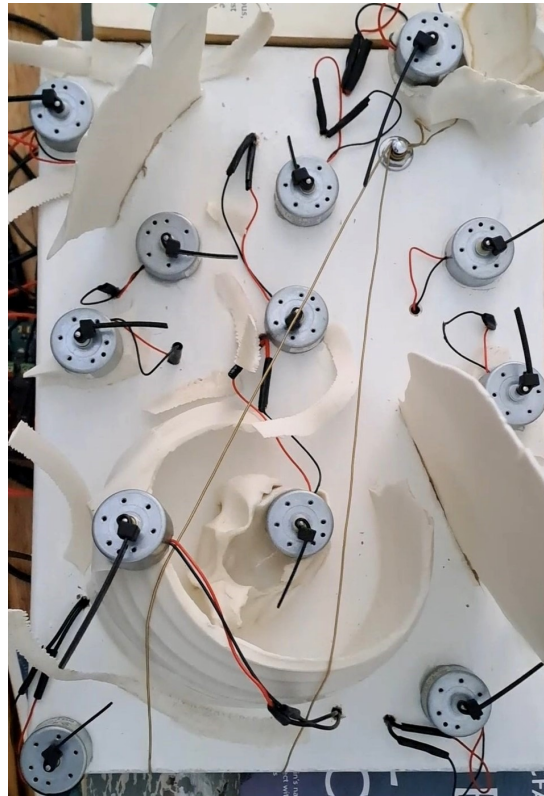


Figure 6.4: *Porcelain Music Box* device

short but random time difference

- **Start button:** start the main program, run for twenty-five turns
- **A button:** turn a single random motor softly
- **B button:** turn a single random motor loudly

Portfolio Example 6.1 — *Porcelain Music Box*: Video^a

Video of the work ► shows the dialogue-act generative process (from 0:00) and the other interactive controls (from 01:24). The work was later displayed at Morrab Library in Penzance.

^a. Portfolio/Chapter 6 - Porcelain Music Box/Porcelain Music Box.mov

6.2.6 Evaluation

This evaluation traces the translation of data from the linguistic structure to the mapping logic, physical realisation, and the function of the interface.

Linguistic Structure to Mapping

The LSTM output followed the general structure of the MRDA corpus (Figure 6.5), which kept the device's output tied to conversational patterns rather than a random distribution. On a small scale, the model tended to repeat exact five-token phrases from the original data. However, as these sequences grew longer, it began to diverge; by the time the sequences reached fifteen tokens, over 80% of the patterns were new. The system essentially uses fragments of the conversational data as building blocks for longer sequences.

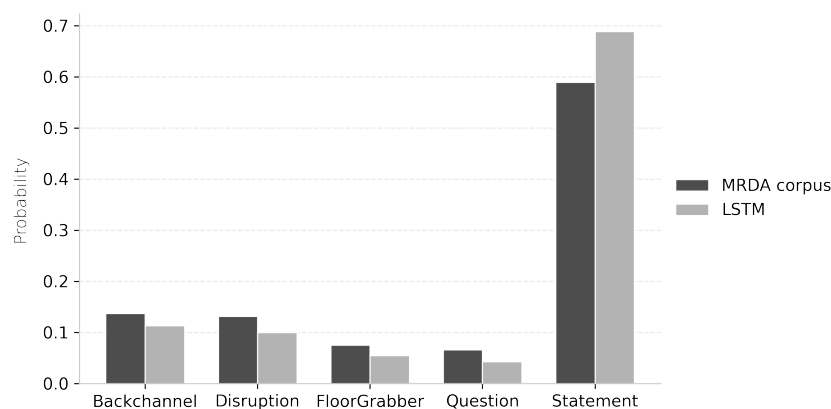


Figure 6.5: MRDA corpus and LSTM distribution similarities

The mapping process translated these sequences into rhythmic patterns of continuation and disruption. To me, this was felt as a compression and stretching of timing (Figure 6.6).

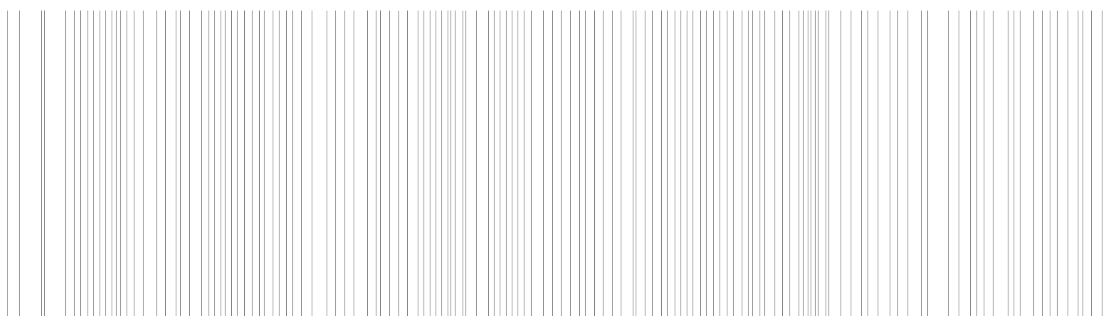


Figure 6.6: Rhythmic onsets from the video recording, showing patterns of dense and sparse material

Mapping to Device

These sequences were realised through the motor-assembly. Unlike the digital model, the physical fragments introduced timbral variation. The irregular shapes and varying thicknesses of the

porcelain meant that identical motor actions produced shifting qualities depending on the strike point.

A core feature of the mapping was the structural mimicry between dialogue acts. For example, the statement-response logic (see `create_statement_response_to_question`) extracts the `root_note_pattern` from a preceding question and loops it. This allowed the responding motor to echo the sequence of the first.

In practice however, it seemed that this mimicry was obscured by a lack of spatial differentiation. Although the software assigned speaker IDs to specific motor pairs to simulate conversational structure, the physical proximity of the motors essentially collapsed this distinction. Rather than a dialogue between distinct points in space, to me, the result felt more like the expression of a single player.

Interface

The USB gamepad provided a reliable method for control, allowing the piece to function as a stand-alone installation where an audience member could trigger new sequences. The manual controls, such as the directional sweeps, added some variety to the output via direct control.

In this work, the source material and mapping process were opaque. While a screen or text explanation could clarify the link between conversational patterns and motor actions, my purpose with the device was to produce abstracted conversational patterns rather than a literal display of linguistic data.

6.2.7 Summary

In *Porcelain Music Box* I developed a method for translating conversational structure into musical material. By training a text model (LSTM) on the MRDA corpus of dialogue acts, I explored linguistic labels as an abstracted source for sequence generation. This approach allowed the realisation of multi-speaker, non-verbal conversations among the device's motors and porcelain fragments. The mapping process translated the dialogue acts into patterns of continuation and disruption, a metaphorical interpretation of conversational flow, to control the timing and dynamics of the percussive events.

So how audible is this as a linguistic structure? In short, not very. At least, I would argue that the metaphorical interpretation of the linguistic structure—the patterns of continuation

and disruption—can be heard, but the impression of any of this as a literal conversation relies on an understanding of the work’s process that is not revealed in the work itself. To me, this does not matter very much; my interest is in the compositional process of working with this dataset, with these objects and microcontrollers, and attending to the results. As I find it, there is an enjoyment and playfulness to the process, and while the aesthetic result matters greatly to me, the visibility of the underlying logic does not. I develop contrasting approaches in later works, with *Seven Studies* using a similarly opaque strategy, while *Word games* makes the connection between text, mapping process, and sound explicit.

6.3 Seven Studies for Electric Motors

6.3.1 Overview

The second work in this chapter, *Seven Studies for Electric Motors* addresses the creative affordances and limitations of text models and their application to linguistic speech structures, by embedding a small language model to generate syntax sequences. The generated sequences are mapped to motors and found objects, exploring the strategies for improvisational interaction that result. My approach to improvisation with these sequences was prompted in part by methods for the electronic manipulation of speech described in Chapter 2—and in this creative project I aimed to apply looping, speed, amplitude and other manipulations in real-time to generated syntax sequences, rather than recorded speech audio.

The work is built around dependency relations; structures that describe syntactic connections between words. Here, I draw on Barlow's approach in *Prog  thal Percussionian*, which computationally mapped parts of speech to a musical score for six percussionists, framed as a course in this new language.¹¹ The key difference is the embedding of a small language model for text generation and live improvisation, rather than a scored work. The dependency relation labels are mapped to sequences of motor and light activity. This algorithmic process of text generation, parsing, and mapping provides a means to create musical sequences that follow linguistic patterns, sequences that would be impractical to compose or realise manually. Found objects and actuators—springs, copper foil, sticks, paper, plastic, broken terracotta—are attached magnetically and rearranged between each study, introducing a modular approach to found materials.

Objectives

1. **Linguistic Structure:** Generate text and extract dependency relations.
2. **Mapping:** Determine motor movements and light patterns through dependency relation labels.
3. **Device:** Build a portable modular system using motors, LEDs, and magnetically-attached found objects.

11. Barlow, *Prog  thal Percussionian for the Advanced Beginners*.

4. **Improvisation:** Develop an interface to manipulate the sequences in real-time, while physically swapping objects to reshape the sound.

6.3.2 Linguistic Structure

An objective for *Seven Studies* was to create a self-contained, portable system, so that I could travel and perform with the device in different locations. This required embedding the language generation process; the computational requirements of large language models make them unsuitable for this kind of local deployment,¹² and the need for a stable internet connection would limit venues for performance. The selection of a specific small language model—Smollm2:135m¹³—was led by this constraint. While small models often lack the accuracy of their larger counterparts,¹⁴ their value in this project lies in their ability to efficiently generate novel, grammatically coherent sentences.

In practice, the model was prompted with the instruction: ‘Generate a sentence about $\mathbf{x}$ ’, where $\mathbf{x}$ is a random word (supplied by the Wonderwords¹⁵ library). Each sentence was then parsed to extract the sequence of syntactic dependency relations (using spaCy¹⁶), which form the input for the mapping process.

These dependencies describe the relationship between words in a sentence.¹⁷ For example, the following sentence was generated by the model: *Mangos are juicy fruits that come in various colors and flavors*. Converted to dependencies, this returns the sequence: `nsubj, ROOT, amod, attr, nsubj, relcl, prep, amod, pobj, cc, conj, punct`. At the centre of this work is the use of these labels as the basis for organising motor and light sequences.

12. Fali Wang et al., ‘A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness’, *ACM Transactions on Intelligent Systems and Technology* 16, no. 6 (2025): 1–87, <https://doi.org/10.1145/3768165>.

13. Loubna Ben Allal et al., ‘SmolLM2: When Smol Goes Big—Data-Centric Training of a Small Language Model’, *arXiv preprint arXiv:2502.02737*, 2025, <https://doi.org/10.48550/arXiv.2502.02737>.

14. Wang et al., ‘A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness’.

15. Maxim Rebguns, *Wonderwords*, 2020, accessed 25 February 2026, <https://github.com/mrmaxguns/wonderwordsmodule>.

16. Matthew Honnibal et al., ‘spaCy: Industrial-strength Natural Language Processing in Python’, 2020, <https://doi.org/10.5281/zenodo.1212303>.

17. Marie-Catherine De Marneffe et al., ‘Universal Stanford dependencies: A cross-linguistic typology.’, in *LREC*, vol. 14 (2014), 4585–4592, accessed 25 February 2026, http://www.lrec-conf.org/proceedings/lrec2014/pdf/1062_Paper.pdf.

6.3.3 Mapping

The mapping process involved converting the generated dependency relations into motion, sound, and light (Figure 6.7). This system was designed around an analogy: the machine-generated syntax provided the grammar to structure events in time, while fragmented approximations of human speech prosody provided the vocabulary for each event. To create this vocabulary, audio recordings of speech were analysed for pitch and amplitude, then translated into a set of over 250,000 motor activation grids, each representing a distinct sequence (Figure 6.8).

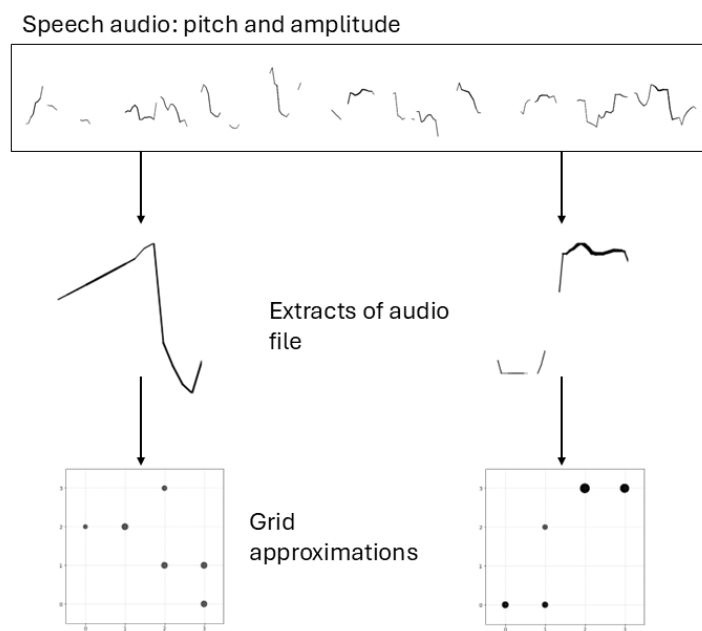


Figure 6.7: *Seven Studies* audio from speech recordings analysed for pitch and amplitude, translated into a 4x4 grid where point size represents motor speed.

This was a two-stage process: each speech audio file was represented as an F0 pitch and amplitude curve, and segmented on silence. Each curve was then reduced to a sequence of points in a grid, where x represents time, y represents F0 pitch estimation, and point size corresponds to average amplitude. This extreme simplification is related to Ablinger's grid-based *Quadraturen*¹⁸ method, applied here to generate a broad vocabulary of motor and light patterns. In performance, a randomised set of individual patterns correspond to dependency relation labels, persisting until a remapping of the vocabulary (see 6.3.5).

This translation of sentence to motor sequence functions in the work as a hidden process. My

18. Peter Ablinger, *Quadraturen*, 1995 -, accessed 25 February 2026, <https://ablinger.mur.at/docu11.html>.

purpose was not for the audience to consciously decipher the linguistic source, but to experience the material results of this process. To make the resulting patterns more easily discernible, I mapped the brightness of each LED directly to the speed of the adjacent motor. Jensenius et al. describe these as *sound-accompanying gestures*, in tracing the sound-producing gestures.¹⁹

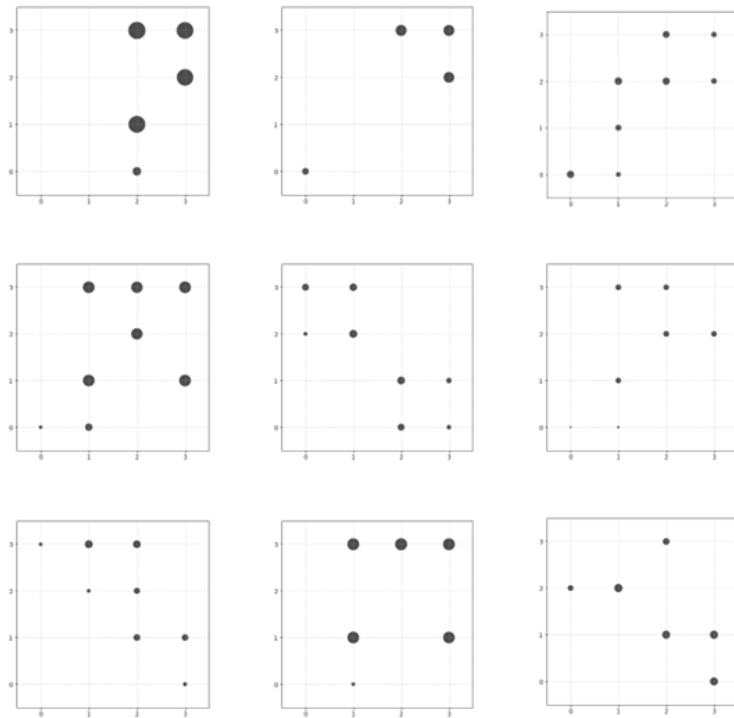


Figure 6.8: *Seven Studies* example grids

6.3.4 Device

I used sixteen motors, of the kind found commonly in CD players. These were also used in *Porcelain Music Box*, and were chosen for their relatively low noise level and ease in vertical mounting. These were mounted on layers of rubber and cork, and held in place with rubberised sealant—all to reduce motor noise. The motors were controlled by a Raspberry Pi, and four DC motor hats (Figure 6.9).

LED's were chosen to mirror the patterns produced by the motors, and make the gestural qualities of the motors clearer. Sixteen LED's were mounted next to each motor, and controlled

19. Alexander Refsum Jensenius and Marcelo M Wanderley, 'Musical gestures: Concepts and methods in research', in *Musical gestures*, ed. Rolf Inge Godøy and Marc Leman (Routledge, 2010), 24–47.

by the Pi via an Arduino and PWM driver.

A perspex panel was used as the surface of the device, on which objects would be placed. This perspex was mounted on speaker spikes and rubber feet to reduce feedback and noise.

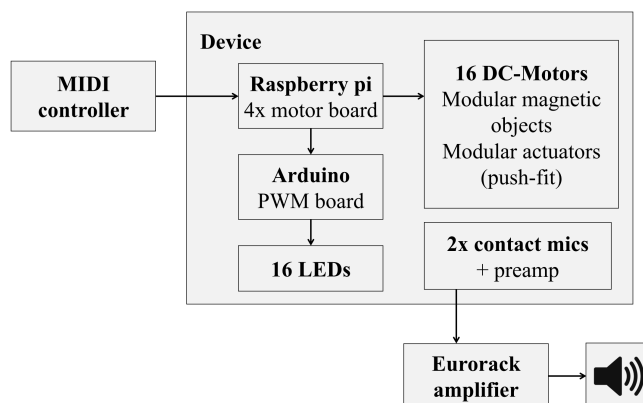


Figure 6.9: *Seven Studies* design

The material approach for *Seven Studies* aligns with the tradition of electromechanical works that celebrate the unpredictable qualities of found materials.²⁰ One limitation of previous works was the fixed objects, and so this uses a modular system instead, allowing objects and actuators to be added, removed, and reconfigured during performance. Each object was mounted on a short wooden cylinder with a magnet at its base, which allowed for quick placement and adjustment. The advantage of this approach was that a wider range of materials could be used, and the material palette included porcelain, drum skin, thick copper foil, sticks, springs, stone, metal bells and bars.

Actuators were similarly modular. I fit each motor with a hexagonal shaft to hold carved wooden tops. These tops acted as mounts for a variety of striking and scraping implements fashioned from found materials including plastic cable ties, springs, plastic bags, and brush (Figure 6.10).

For amplification I used two contact microphones, a preamp and Eurorack amplifier, for high sensitivity and low noise. One limitation of the motor drivers was a PWM control with a maximum frequency of 1.6KHz, which meant that the PWM control was audible. While different drivers were used in the following project to address this, Eurorack filters were used

20. Jon Pigott, 'Across Fields: Sound, art and technology from an electromechanical perspective', *Organised Sound* 22, no. 2 (2017): 276–285, <https://doi.org/10.1017/s1355771817000188>; Andrea Valle, 'Making acoustic computer music: The Rumentarium project', *Organised Sound* 18, no. 3 (2013): 242–254, <https://doi.org/10.1017/s1355771813000216>.

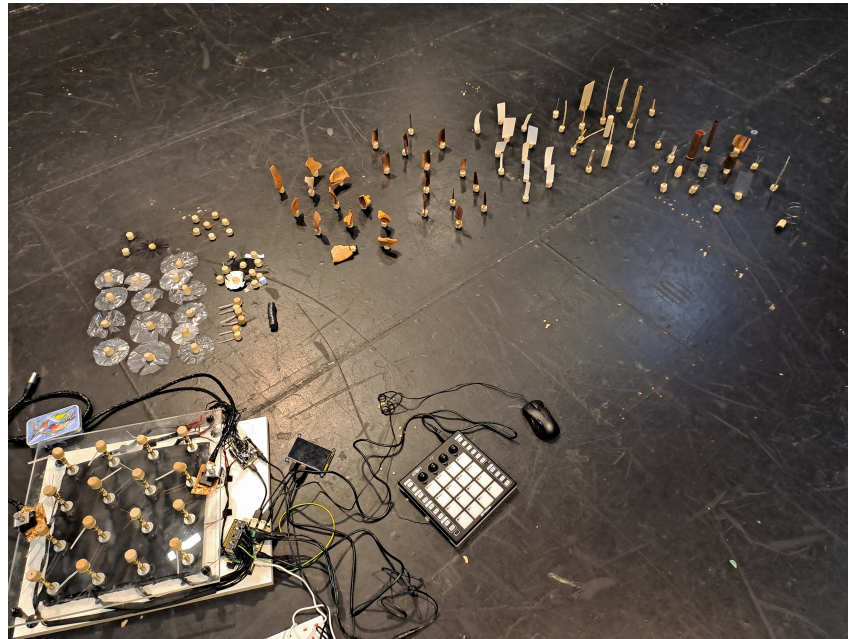


Figure 6.10: *Seven Studies* objects and actuators

here to reduce noise.

6.3.5 Improvisation

Seven Studies was designed as an interaction between a generative system and a human improviser. While the system incorporates an autonomous process, it does not make musical decisions. Instead, the interaction is built around the performer's real-time response to the system's output. Parkinson and Dunning describe this as a machine that plays rather than runs;²¹ this places the performer in a curatorial role, responsible for interpreting and shaping the machine's output in real-time.

A rule-based interface was implemented using an ATOM MIDI controller, chosen for its combination of touch pads and rotary encoders. The 16 LED touch pads were configured as latching switches, allowing individual motors to be muted or unmuted. The four rotary encoders managed global parameters:

- **Overall speed** by multiplying overall durations.
- **Quantisation** allowing for rhythmic variation.
- **Overall velocity** set the mean speed of the motors while preserving local variation.

21. Adam Parkinson and Graham Dunning, 'Improvising with Machines: A taxonomy of musical interactions', *Organised Sound* 29, no. 3 (2024): 340–351, <https://doi.org/10.1017/s1355771824000268>.

- **Rest post-sentence** allowed adjustment of the silence after sentences.
- **Rest post-grid** in combination with a button press introduced silence between each dependency label for textural control.

A final button instantly remapped the dependency labels to new motor patterns. This control scheme established the machine as a sequence generator, offering a continuous stream of motor and light patterns, while in performance I acted as an interpreter, responsible for shaping phrasing, dynamics, tempo, and texture.

Code Example 6.2 — *Seven Studies: Scripts*^a

Three programs were run simultaneously on the device: a python script to generate sentences using the model (via Ollama^b), a Pure Data^c patch to handle timing and controller input, and python scripts to control the motors and lights.

- `generate_sentences.py` for generating sentences. `</>`
- `pd_patch.pd` for coordinating between the small language model, mapping, and forwarding data to motors and LEDs. `</>`
- `OSC_recieve.py` for controlling the motors, and receiving data from the puredata patch. `</>`
- `LED_recieve.py` for controlling the LEDs, and receiving data from the puredata patch. `</>`

^a. Code/Chapter 6 - Seven Studies for Electric Motors/...

^b. Francisco S Marcondes et al., 'Using ollama', in *Natural Language Analytics with Generative Large-Language Models: A Practical Approach with Ollama and Open-Source LLMs* (Springer, 2025), 23–35, https://doi.org/10.1007/978-3-031-76631-2_3.

^c. Miller Puckette et al., 'Pure data: another integrated computer music environment', in *Proceedings of the second intercollege computer music concerts* (1996), 37–41, accessed 25 February 2026, <https://msp.ucsd.edu/Publications/icmc97.pdf>.

The video work, *Seven Studies for Electric Motors*, captures seven configurations of objects and actuators in a real-time performance using the improvisational interface. Controller settings for each track used either a single value or a range (Figure 6.11). I arrived at each position through improvisation with the interface and materials. Texts generated throughout each study were not saved: the musical patterns were defined by the grammar, and linguistic meaning was lost in the process.

Track	Speed	Amplitude	Note length	Quantization	Rest post-grid	Rest post-sentence
1						
2						
3						
4						
5						
6						
7						

Figure 6.11: *Seven Studies* controller settings

Portfolio Example 6.2 — *Seven Studies*: Video

Videos of each study can be seen in the *Portfolio*.^a Unlike live performance, where the objects were often changed gradually, each setup was changed between takes.

- 1 copper, springs ►
- 2 wood, springs, copper ►
- 3 plastic, paper ►
- 4 copper, springs ►
- 5 terracotta, paper, wood ►
- 6 springs, copper, paper ►
- 7 copper, springs ►

^a. Portfolio/Chapter 6 - Seven Studies for Electric Motors/...

6.3.6 Evaluation

The following evaluation examines the translation of linguistic structure to mapping, the mapping to physical found objects, and the interaction between performer and generative stream.

Linguistic Structure to Mapping

I found that the primary affordance of embedding the language model was the ability to generate novel musical material that adhered to an internal logic, in a self-contained and portable device. The trade-off was a limitation in generative complexity compared to larger models, but

even the small model could produce syntactically rich structures sufficient for musical sequence generation. The choice to generate text live, rather than drawing from a static corpus, prioritised an in-the-moment unpredictability which was central to the work's character. However, one limitation of the small model was the lack of control over sentence length, a feature often present in larger models (to a lesser extent) given the encoding of text.²²

The speech audio-derived sequences were broad enough to create variety within the constraints of the 4x4 motor grid. The scale of the 250,000-grid vocabulary provides evidence for this variety; the grids range from low-speed motor activations to dense, higher-velocity patterns (as illustrated in the earlier range of Figure 6.8).

Mapping to Device

My approach to repurposing a language model and dependency parser for musical generation aligned with the repurposing of materials. Because the translation from syntax to sound was a hidden process, the audience's attention was directed away from the abstract grammar and towards its physical enactment. While the objects themselves had no specific conceptual significance, the aesthetic of reuse was a key feature. I chose objects for their sonic qualities, picked from materials to-hand and kept or discarded through improvisation with the device. A broad aim was to create a varied palette of noise-based sounds, from the pitched qualities of metal springs to the dry clicking of paper and plastic. Another was to explore variation within each material; objects repeated in different sizes—similar pieces of broken terracotta, copper sheet of varying dimensions.

Variety in the objects meant a higher degree of sonic differentiation within the sequences, but changing these objects broke the continuity of the patterns. Another factor I found obscured the sequences was the speed at which the machine operated. Study 4 (*copper springs ii*) shows an example of the rest between each pattern set to a longer duration. Here, we can hear each distinct token, with the motor order and relative amplitude set by the generated dependency relations. In this case, moments of repetition were clearer than in continuous uninterrupted sequences, although I found that by listening to the machine for extended periods a kind of logic emerged from the recurring patterns, in terms of which pattern was likely to follow from another.

22. Tairan Fu et al., 'Why Do Large Language Models (LLMs) Struggle to Count Letters?', *arXiv preprint arXiv:2412.18626*, 2024, <https://doi.org/10.48550/arXiv.2412.18626>.

One limitation of the modular approach to found objects was the time required to prepare a new object: five minutes to glue the material and attach magnets. While the system allowed for new objects to be tested and integrated, this preparation time obstructed an improvisatory approach to object selection. This is specifically addressed in the following work (*Word games*).

Improvisation and Interface

My improvisational experience was characterised by the ability to shape and respond to this generative stream in real-time, and while the unpredictability of the output prompted the use of certain objects or actuators in response, these objects and actuators did not influence the machine's behaviour. The device could appear autonomous, generating a continuous stream of novel material without direct control. On the other hand, the rearranging of physical objects and occasional manipulations via a MIDI controller clearly shaped the musical qualities of the output.

The MIDI interface provided high-level influence over the execution of sequences, while the physical rearrangement of objects functioned as a form of acoustic remapping. This allowed for shifts in timbral character without altering the underlying pattern. In performance, it was left to the audience to interpret the precise nature of the human-machine interaction. However, the one-way control contrasts with the reciprocal interaction seen in several surveyed works through Chapter 3, and could be addressed in future by adding audio or visual input back into the system.

6.3.7 Summary

Through this project I have demonstrated one possibility in the application of a small language model to generate musical-linguistic sequences, where an interface, modular objects, and actuators allow the improvisational shaping of a generative output.

The following work, *Word games*, combines a fixed text with the improvisatory modular mechanical approach developed here, through an exploration of word embeddings as the basis for musical material.

6.4 Word games

6.4.1 Overview

In this final work I address R3 through the use of word embeddings as the basis for a generative musical system. This is the third linguistic feature of this chapter, each of which are a response to the relative lack of existing works exploring grammar structures (Chapter 2). The work itself is a ten-minute piece for a DIY device that translates words into patterns of light and sound, through the use of word embeddings. I structure the work as a self-referential explanation of its own process.

This approach differs from both *Porcelain Music Box* and *Seven Studies* in that I make the link between text and sound explicit. However, a common aim is the use of linguistic speech structures as the basis for composition, and I draw again on Barlow's *Progéthal Percussian*²³ as an example of a translation from text to abstract sound. In the material aspects of the work, the modular approach to found objects is developed in response to the challenges of *Seven Studies*.

Objectives

1. **Linguistic Structure:** Take word embeddings as the basis for a generative system, reducing to a manageable number of dimensions.
2. **Mapping:** Define a process of translation from reduced embeddings to musical parameters.
3. **Device:** Build a DIY device with an improved approach to modular found objects.
4. **Composition:** Write a text to drive the device, as a fixed composition.

6.4.2 Linguistic Structure

Word embeddings represent words as vectors in a multi-dimensional space.²⁴ These capture semantic relationships, where similar words cluster together—for example, ‘colours’ share similar vectors. My aim in this work was to highlight the properties of word embeddings through sound, in which semantically similar words would produce similar sequences.

23. Barlow, *Progéthal Percussian for the Advanced Beginners*.

24. Felipe Almeida and Geraldo Xexéo, ‘Word embeddings: A survey’, *arXiv preprint arXiv:1901.09069*, 2019, <https://doi.org/10.48550/arXiv.1901.09069>.

Standard word embeddings typically range between 100 and 300 dimensions, with recent large language models using embeddings of over 1,000 dimensions.²⁵ Attempting to represent a high-dimensional embedding through nine motors proved impractical, and so I used the GloVe 50-dimension embeddings²⁶ as a foundation. I reduced this dataset significantly to seven dimensions using Uniform Manifold Approximation and Projection²⁷ (UMAP) and filtered to the 100,000 most frequent words to reduce outliers.

Dimensionality reduction techniques like UMAP prioritise local structure, so relative distances between distant clusters may be distorted. In the context of this work, my priority was local clustering (i.e., ensuring semantic groups remained intact) and so some distortion of global relationships was an acceptable trade-off to achieve meaningful clusters. This reduction process was iterative, involving multiple rounds of parameter tuning. I validated the results intuitively by querying words with known synonyms to ensure the resulting clusters were semantically plausible.

Following reduction and filtering, the vectors were mean-centered to align the distribution around zero. The minimum and maximum values for each of the seven dimensions were calculated to establish the boundaries of the space. The data was then normalised, and the 5th and 95th percentiles were flattened. This emphasised differences within the centre of the distribution—the area containing the most commonly-used words—ensuring that while synonyms would sound similar, they would not be identical (Figure 6.12).

6.4.3 Mapping

The mapping from word embedding to motors was achieved through a custom Python script. In this system I convert the 7-dimensional vectors described under *Speech Sources* into instructions for the hardware. To generate a motor sequence, the system retrieved the vector for a specific word. These seven values acted as interpolation indices to generate short sequences of motor speeds and durations (and corresponding LED sequences). Dimensions 0, 5, and 6 interpolated between lists of discrete sequences, while dimensions 3 and 4 interpolated between curves.

25. Roberto Boselli, Simone D’Amico and Navid Nobani, ‘eXplainable AI for word embeddings: a survey’, *Cognitive Computation* 17, no. 1 (2025): 19, <https://doi.org/10.1007/s12559-024-10373-2>.

26. Jeffrey Pennington, Richard Socher and Christopher D. Manning, ‘GloVe: Global Vectors for Word Representation’, in *Empirical Methods in Natural Language Processing (EMNLP)* (2014), 1532–1543, accessed 25 February 2026, <http://aclweb.org/anthology/D14-1162>.

27. Leland McInnes, John Healy and James Melville, ‘Umap: Uniform manifold approximation and projection for dimension reduction’, *arXiv preprint arXiv:1802.03426*, 2018, <https://doi.org/10.48550/arXiv.1802.03426>.

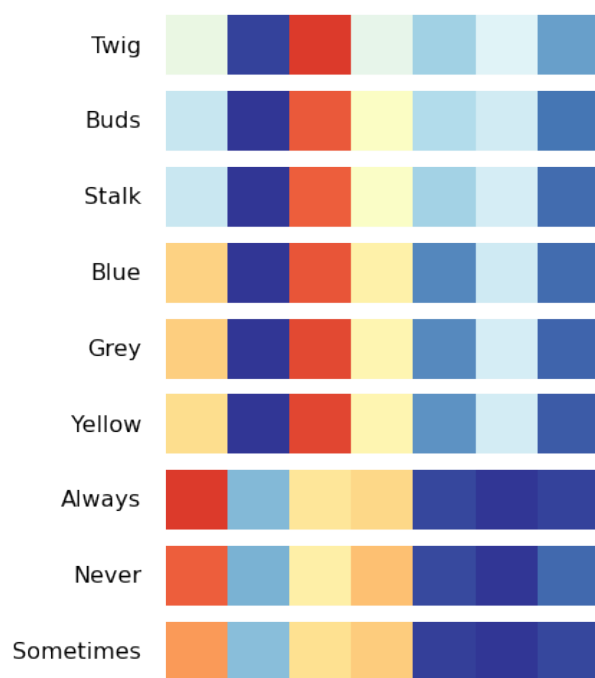


Figure 6.12: Example of *Word games* 7-dimensional embeddings, visualised using coloured squares to show similarities between related words

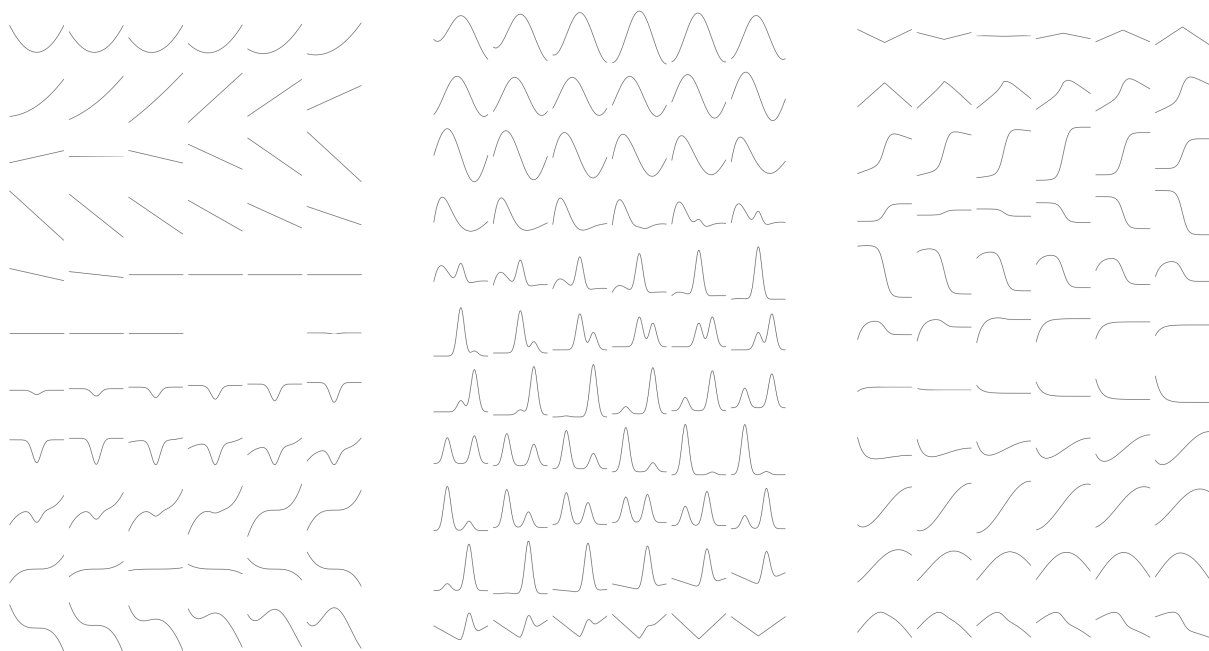


Figure 6.13: *Word games* curve interpolation for a single embedding dimension; the initial set of curves were chosen manually.

Both sequences and curves were determined by hand, to define the possible space of the output. This use of both string and curve interpolation allowed a single dimension to control a complex parameter, meaning that semantically related words could share similar patterns,

with nearby words showing slight deviations from the same pattern. Examples of the curve interpolation can be seen in Figure 6.13, where a single value between 0 and 1 returns a range of curves. String sequences were used for motor sequence order, as in the following example which is a short extract of an interpolation, producing a list of related motor sequences:

```
[0, 1] [0, 1] [0, 1] [0, 1, 5] [0, 1, 5] [0, 2, 5] [0, 2, 5] [8, 2, 5] [8, 2,
5, 5] [8, 2, 5, 5] [8, 2, 8, 5] [8, 7, 8, 5] [8, 7, 8, 5] [4, 7, 8, 5] [4, 7, 8,
5] [4, 7, 8] [4, 7] [4, 7] [4, 7] [6, 7] [6, 7] [6, 7, 2] [6, 7, 2] [6, 8, 2]
```

Code Example 6.3 — *Word games: Mapping*

One challenge was developing a method for producing consistent sequences from a single word, so that a word would have a persistent associated sequence without relying on stochastic distributions (i.e., the same word would always produce the same sequence). To achieve this, the specific mapping was as follows:^a

- **Dimension 0—Motor path:** This determines the physical path of the pattern across the nine motors. The value interpolates between pre-defined sequences of motor indices, for example blending a repetition [0, 1, 0, 1...] with a scattered group [8, 2, 5, 3...].
- **Dimension 1—Duration:** This value maps linearly to the total duration of the phrase, ranging from 0.3s to 1.7s.
- **Dimension 2—Density:** This determines the overall number of notes within the phrase, quantised to an integer between 2 and 6.
- **Dimension 3—Rhythmic distribution:** This selects the onset times. The value interpolates between a library of ‘density curves’ (e.g., Gaussian, sinusoidal, step).
- **Dimension 4—Dynamic contour:** Similar to the rhythmic distribution, this interpolates between curves to define the amplitude envelope of the phrase.
- **Dimension 5—Articulation:** The system interpolates between sequences of articulation modes, such as staccato (‘S’, 0.1s), legato (‘L’, full duration), or staccatissimo (‘Ss’, 0.05s).
- **Dimension 6—Envelope Nuance:** This interpolates between sequences of individual note envelopes, such as ‘Fade-In’ or ‘Flat’.

For synchronisation between light and sound, the system pre-computes a complete event

list for each word. This function calculates specific onset times, motor velocities, and LED brightness, applying a correction to the amplitude to better match perceptual hearing.

a. Code/Chapter 6 - Wordgames code/words5.py

6.4.4 Device

I assembled the device from found objects, nine electric motors, and LEDs. A camera was mounted on the device, and text overlaid onto the video feed. Kanta Horio's *Particle*²⁸ informed my presentation of the work, in displaying the motors, objects, and the hands of the performer from above. Objects and actuators were replaced throughout, making each performance slightly different.

The device used a Raspberry Pi as the main controller, with two Arduino microcontrollers for motors and LEDs (Figure 6.14). The Pi was also used to overlay text subtitles onto the live feed from a camera (Figure 6.15). Amplification used exactly the same system as *Seven Studies*—a preamp, Eurorack amplifier and two contact microphones.

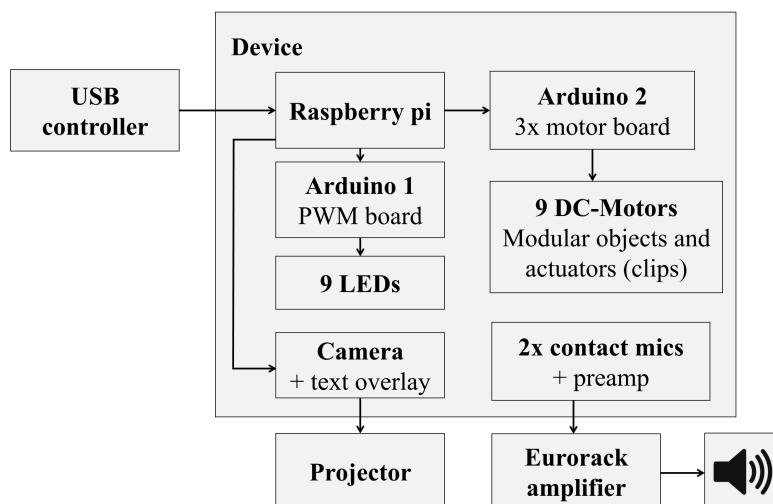


Figure 6.14: *Word games* design

One of the main problems with the modular approach in *Seven Studies* was the magnets. Powerful enough to hold an object as it was struck, these were prone to sticking together when travelling with the device. In *Word games*, I used spring clips instead to allow the replacement of objects during performance: small foldback clips for the actuators, and large bulldog clips for

28. Kanta Horio, *Particle*, 2008, accessed 25 February 2026, <https://horiokanta.com/particle/>.



Figure 6.15: *Word games* device with camera

the objects.

The second problem with the approach in *Seven Studies* was that making new objects took at least five minutes. That process obstructed an improvisatory, experimental approach to found sounds. For *Word games*, I created objects by supergluing string, material, wire and others between two pieces of soft foam. This meant that new objects could be created in a matter of seconds, and allowed more experimentation with a wider variety of materials.

As the purpose of the device was to play a predefined text score, the interface was minimal. This consisted of a small USB controller with a rotary encoder, and six buttons. One button was used to play and pause the output, with the encoder to select sections of the work, essential for rehearsing with the device.

There was no other user input during performance. Instead, the work presents the pre-composed sequences, modified by the improvisatory placing and replacing of objects during the piece.

6.4.5 Composition

I wrote the composition as a text score for the device to play back through motors and LEDs, with subtitles overlaid onto the camera feed. This text describes the method: information about word embeddings, how they are processed by the machine, how features are mapped to the motors, and examples of semantically similar words and their associated sounds. My intent was for the text to function as an embedded essay, in parallel with the machine's reconstruction of the same text.

I used repetition of single words to emphasise the musical aspects of the resulting patterns, for example (*word, delay*):

```
("This machine turns", 0.05),  
("words", 0.2),  
("words", 0.2),  
("words", 0.2),  
("words", 0.2),  
("words", 0.2),  
("words into patterns", 0.05),  
("of light", 0.1), ("and sound", 0.05),  
("sound", 0.3),  
("sound", 0.3),  
("sound", 0.3),  
("sound.", 3.6)
```

In later sections of the work I present groups of words in a shared semantic field—with the aim that the audience would attend to the resulting patterns more than the text itself:

```
("AND", 0.05),  
("WITH", 0.05),  
("A", 0.05),  
("BUT", 0.05),  
("THEY", 0.05),  
("THEM", 0.05),  
("THEIR", 0.05),
```

("SCREEN", 0.1),
("VIEWING", 0.1),
("PICTURES", 0.1),
("PHOTOGRAPH", 0.1),
("DISPLAY", 0.1),
("IMAGES", 0.1)

My choice of particular words was led by experimentation with a separate script, which played a typed word through the device. In composing the text, I replaced certain words if the resulting pattern did not work musically. For example, in the ‘semantic fields’ section of the work, the grouped patterns were intended to contrast, and so the final text often reflects musical choices rather than linguistic.

Portfolio Example 6.3 — *Word games: Video*

The *Portfolio* contains a video of the work.^a ► My aim with the work was for it to be self-explanatory, and so some aspects of the musical choices are described below.

1. **01:53** colours were chosen here, rather than possible alternatives, because they showed some small differences while being very closely related—an alternative group of ‘fruits’ was deleted in a later draft, as they sounded identical.
2. **02:53** shows repetition of the word ASPECT. In general, repeated words are not connectives (AND, THE, OF etc.) as they sound very similar to other connectives, and repeating more unusual patterns was a way to highlight variety in the output.
3. **06:04** shows an example of words which may not be expected to sound the same, rendering as identical (SET and OF).
4. **06:38** presents a list of letters, chosen to provide some rhythmic contrast as they were rendered relatively dense compared to other words.
5. **06:55** double-digit numbers sound surprisingly different to the three-digit numbers of **08:05**, which is the main reason for their inclusion.
6. **08:16** increases the rest between words, simply to create some contrast between the different groups here.

a. Portfolio/Chapter 6 - Word games/wordgames.mov

6.4.6 Evaluation

The following section considers each stages of the translation process, from linguistic structure to mapping, to device, and finally the compositional approach.

Linguistic Structure to Mapping

I found that using word embeddings allowed a semantic exploration of the device. While the potential space for the mapping was defined, in my experience there was enough variety within the vocabulary—and a large enough vocabulary—to discover new material by typing words into the device. In this way, composing for the device simply involved editing a text, and listening to the results.

Instead of controlling parameters directly, the seven embedding dimensions served as interpolation indices. This is key because dimensions were used to blend between discrete sequences (for motor order or articulation) and curves (for rhythmic density and amplitude contours). This curve interpolation allowed a single continuous value from the embedding to control a complex musical parameter. For instance, interpolating between different rhythmic curves ensured that closely related group of words shared a similar underlying rhythmic character, with the degree of variation derived from their exact vector positions.

I limited the initial dataset to the 100,000 most common words from the GloVe corpus, which excluded many terms, including the plural ‘embeddings’, which could have been used in the composition. A related issue was the lack of consideration for punctuation and the presence of unusual tokens in the vocabulary (like 10:4), which were never explored because I chose to strip punctuation before playback. Future iterations could remove these obscure words, which would extend the expressive space of the common words. However, a balance to ensure semantically similar words sounded similar was part of the fine-tuning of the existing system. The possibility of adding more mapping dimensions could be explored to increase the complexity and variety of the musical output. Finally, the system could involve audience participation in typing words during performance, if a broader vocabulary (or alternatively sub-word embeddings) were used.

Mapping to Device

Following *Seven Studies*, in this work I developed the modular approach to found objects. In *Word games* I used small spring clips to hold objects and actuators in place—and instead of the

five minutes it took me to make an object in *Seven Studies*, I could test new objects here in seconds, which supported an improvisatory approach to the materials of the work (Figure 6.16).

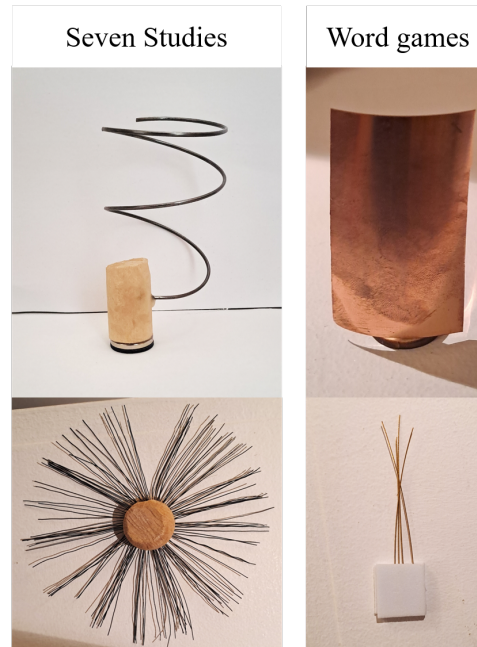


Figure 6.16: Relative simplicity of *Word games* objects and actuators to *Seven Studies*

The materials and arrangement of objects and actuators defined the sound. This idea was explored in Chapter 4, with the difference here the ability to easily change the objects. I felt that the variable loops created by the device sustained more interest than direct repetition. This of course worked against a consistency in the realisation of the word embeddings, and so the object-actuator arrangements were usually simple constructions involving a spinning flexible actuator and fixed object, rather than more unpredictable arrangements like the bells and band-suspended ball bearing of *Motorcase I*, to avoid a level of disruption that would obscure the patterns beyond all recognition.

Composition

The composition was a self-referential text score. It explains its own process while the machine translates it into light and sound, using word repetition and semantic fields to highlight the resulting musical patterns. In this sense, it does what I set out to do.

However, there are several limitations with word embeddings (of the particular kind used here) which I did not address in the piece. Word-level embeddings means that homographs share an embedding, so for example, the word *bank* is represented as somewhere between a bank

robber, a river bank, a bank of clouds, and an aircraft turning. A second limitation is bias in word embeddings, reflective of wider societal biases.

So why were these limitations not addressed? The homographs example is simply not very interesting to listen to, as repetition of single words was already a central aspect of the piece. And while algorithmic bias is a critical limitation,²⁹ these issues were not reflected within the reduced seven-dimensional space of the work. For instance, the well-documented gender biases found in high-dimensional word embeddings, such as the association between ‘homemaker’ and ‘woman’, could not be reproduced. Furthermore, the vector arithmetic popularised by examples like `king - man + woman = queen` did not work within the mapping process. In this work, such operations involve modulating interdependent parameters, such as the simultaneous shifting of rhythmic density and duration. A listener may perceive that a sound has changed, but they cannot gauge the specific direction or magnitude of the change, the *how much* of the operation. To address this, consistent and independent spaces would be needed for each dimension—for example rhythmic density (say, dense to sparse) and timbre (dull to bright)—and in this case, moving from `dense + high` along the dimension corresponding to density, to `sparse + high` would be perceived as a shift along a single dimension, with a sense of the magnitude of the shift.

6.4.7 Summary

In *Word games* I addressed R3 by creating a generative musical system based on word embeddings. The work demonstrates a direct link between linguistic meaning and abstract sonic output, in contrast to the hidden relationship in *Porcelain Music Box* and *Seven Studies*.

The evaluation shows that while vector arithmetic was musically unsuccessful, the use of embedding dimensions as interpolation indices for rhythmic and amplitude curves was effective in the compositional context of the work. In the material aspects of the piece, the shift toward a modular physical interface with spring-clip actuators supported an improvisatory approach to found objects.

In summary, the work demonstrates the viability of using dimensionality-reduced word embeddings as a generative tool for composition. The practical and expressive considerations raised

29. Orestis Papakyriakopoulos et al., ‘Bias in word embeddings’, in *Proceedings of the 2020 conference on fairness, accountability, and transparency* (2020), 446–457, <https://doi.org/10.1145/3351095.3372843>.

by this and other projects in this chapter are addressed in the following discussion.

6.5 Discussion

In this discussion I evaluate how linguistic structures were mapped to mechanical action across the three projects. This addresses the translation of syntax to sound, the constraints of the physical found objects, and the resulting interaction between performer and generative stream.

6.5.1 Mapping Linguistic Structure

The structure of each source informed my musical approach. For example, dialogue acts as the more global feature—with each token describing an utterance—suggested longer sequences in *Porcelain Music Box* than the word-level embeddings for *Word games*. Likewise, the discrete tokens of each feature were mapped to discrete sound-events. While nothing in the data demanded this approach, the purpose of each piece was to enact linguistic patterns, reflecting something of the original source material (Table 6.2).

Features were pre-processed to suit the mapping procedures for each work. In all cases, this involved a loss of information. The most extreme example was the reduction from 50-dimensional word embeddings to 7-dimensions, in *Word games*. This involved a trade-off between the number of perceptual musical parameters within the limited capabilities of the motors and objects, and the approximation of the embedding space.

This loss of information is however present in other works. In *Porcelain Music Box* I filtered the dataset to conversations of up to six participants, and using only the highest level descriptors of conversation acts. Likewise, in *Seven Studies*, the dependency relations used were not the most detailed option, and the use of higher-level descriptors was to highlight repetition in the generated structures. The material itself, grid-approximations of speech audio, further demonstrate a low-resolution approach to source material throughout, necessitated by the mapping approaches that followed.

This allowed the works to focus on language-derived sequences. While the audio features or generated audio of previous chapters was arranged into loops or composed into patterns to create some musical structure, here, the existing organisation defines it.

Table 6.2: Comparison of Linguistic Mapping and Sequence Generation

	Porcelain Music Box	Seven Studies	Word games
Linguistic Structure	Dialogue acts	Dependency relations	Word embeddings
Linguistic Resolution	Conversational turns	Word-level grammatical relationships	Individual words
Sequence Driver	LSTM generative model	Small language model	Fixed text score
Mapping Mechanism	Fixed category-based algorithmic generation	Stochastic category-based pattern assignment	Multi-dimensional interpolation

The dialogue acts in *Porcelain Music Box* were extremely limited in terms of information, having only a speaker (from zero to five) and one of six broad categories of dialogue acts. The approach to mapping involved the act category itself as a marker for a set of generative rules. Likewise, the stochastic assignment of speech-derived grid patterns to dependency relations in *Seven Studies* superimposed these patterns onto linguistic structure. In both cases, my feeling was that the musical result could more or less be predicted from the rules and source material. *Word games* seemed to be different in this respect: the generated patterns were still within defined bounds, as with *Porcelain Music Box*, but the generative space was much broader. Because the musical material was generated using several parameters, I could not predict the resulting sequence to the same degree, and this meant that the device produced unexpected results. The trade-off for these unexpected results was that while local semantic similarities were preserved (blue and green sounded similar, blue and never sounded different), the magnitude of the differences were not intuitively perceived (how far is blue from never?).

6.5.2 Material Constraints

My approach to mapping was also informed by the capabilities of each instrument. With *Porcelain Music Box*, the use of timbrally similar objects meant variation was largely derived from rhythm and motor speed, rather than the objects themselves. At least, the variation between objects was subtle, and so the mapping was designed around fluctuating waves of rhythm and amplitude. In both other works, the range of objects meant that the order of motor events was key to the realisation of the linguistic patterns. In *Word games*, this order was determined by the embedding dimension with the most variation. In *Seven Studies*, while motors could be muted,

the order of events was least-altered through the improvisation interface, which concentrated instead on modifying the amplitude, playback speed, quantisation, and other features. Against this order is the modular approach to object selection. In the recording for *Seven Studies*, objects were unchanged throughout each movement, but when improvising live these were often changed while playing. Likewise, *Word games* involved the changing of objects during playback. In both cases, I saw this as a disruption of the underlying structure of each work, and allowed a balance between familiarity and novelty in performance.

6.5.3 Text Models and Agency

On a technical level, the models were very different. The LSTM of *Porcelain Music Box* is a very basic text generator, but suited the task of producing a small set of tokens with similar statistical properties as the dataset. The small language model used in *Seven Studies* was trained for natural language generation, and repurposed in this work for mapping to dependency relations. Likewise, the process for identifying dependency relations also used a pretrained model, Spacy's dependency parser.³⁰ *Word games* instead used an existing set of word embeddings as the starting point, with the text score defining the structure of the work, and modified through the changing of objects in performance. The shared characteristics of each work was the live generation of musical material, derived from the linguistic patterns of speech and text. This live generation meant that new patterns could be generated with the same internal logic.

Within this live generation, time was a primary constraint, with the speed of each model a limiting factor in the responsiveness of the device. In *Porcelain Music Box* this was less of a concern, as a single token generated a several musical events, and the model itself was quick to generate the next token. In *Seven Studies*, this was addressed by generating sentences rather than individual words, and allowing the device to loop or rest while the next sentence was being generated. While *Word games* used a pre-written text, there was a clear artistic appeal in live generation—an intangible in-the-moment musical aspect which added to the works. Future developments in processing speed, model size, and model inference speed will likely make these limitations easier to manage.

Reflecting on the ML-aspect of these works, it is clear to me that the models function as tools for linguistically consistent generation rather than autonomous agents. In this context,

30. Honnibal et al., 'spaCy: Industrial-strength Natural Language Processing in Python'.

the main affordance is ease of use relative to algorithmic approaches. As seen in *Seven Studies*, the machine-personality is not a result of model ‘creativity,’ but an emergent property of the speech-analogy mapping. The device’s level of autonomy remains low, producing material within narrow generative bounds that are constantly modulated by the physical assembly and interface.

6.5.4 Conclusion

This chapter has addressed R3, on the affordances of text models in artistic practice applied to linguistic speech structures. By applying three linguistic features in practice—dialogue acts, dependency relations, and word embeddings—I have shown three possible approaches to generating speech-derived electromechanical music.

The devices in this chapter presented three kinds of performance: an installation, an improvisation, and a pre-written composition. These different settings affected the design of each interface: *Word games* and *Porcelain Music Box* used a minimal interface, relying on the structure of the work to sustain interest. In *Seven Studies*, the interface involved significant modification of the playback, making the playing experience similar to the *Motorcase* devices, in which I shaped the output of the device.

Each work discussed in the chapter was ultimately derived from a linguistic structure, but can the listener tell? I suggested in the summary for *Word games* (6.2.7) that the value of these strategies to me is in the process of developing and attending to a system. In this way, my engagement with these structures is as a tool for creative practice. It might seem that this attitude disregards the listener, but I would argue against this; how the results of each process and their aesthetic qualities are perceived is important to me, only the perception of the underlying linguistic process is not. Would explaining—as Stapleton and Davis explained the technology behind their networked performance³¹—that these pieces are derived from some linguistic pattern, add to a listeners experience? It might, but it also risks frustrating a listener who cannot hear the pattern of dependency relations in *Seven Studies*, and concludes that the work is not worth listening to as a result.

Comparing the works highlights a common challenge: the translation of often high-dimensional linguistic data into DIY acoustic sound production, with all the associated physical constraints.

31. Paul Stapleton and Tom Davis, ‘Ambiguous Devices: Improvisation, agency, touch and feedthrough in distributed music performance’, *Organised Sound* 26, no. 1 (2021): 52–64, <https://doi.org/10.1017/S1355771821000054>.

As noted in the discussion, this process involved a significant loss of information. My creative act was in designing mapping strategies that take advantage of these constraints. By using grid patterns or simplifying data vectors, the hardware limitations shape an aesthetic built on the underlying structure of speech, rather than its sound. While limitations on inference speed imposed constraints on live generation, the modular approach to object selection and parameter mapping expanded the range of possible sound-object interactions. The strategies discussed here have shown how linguistic models can function as an organisational layer for DIY systems, providing a structured yet unpredictable method for navigating the physical constraints of found objects.

Chapter 7

Conclusions

7.1 Introduction

This thesis has presented a practice research project on the applications of computational speech analysis to motor-driven sound art. Through a series of sound-making devices, I have developed three approaches in the portfolio of works: the algorithmic transcription of speech audio, the incorporation of NAS models, and the application of text models to linguistic patterns.

From mapping the contours of a speech signal to generating structure from a text model, the process has involved experimentation with computational methods and creative strategies within the context of DIY practice and live improvisation.

Throughout the project I was guided by three research questions, each addressing a particular computational approach to deriving musical material from speech:

- R1. Which features can be computationally extracted from the speech signal, and how can they be implemented in an electromechanical improvisation and composition practice?
- R2. How can NAS models trained on speech datasets be incorporated into this practice, and what methods can be used to organise generated material?
- R3. What are the affordances and limitations of text models as creative tools, applied to linguistic speech structures?

The following summary addresses the research questions directly. A practice overview is outlined, followed by *Themes* (7.4) which discusses common threads that have emerged; approximation as a generative method, material constraints, interaction, and transparency. The

contributions to knowledge are outlined, including novel interfaces for navigating speech corpora, methods for approximate acoustic resynthesis, and the application of computationally generated linguistic structures to electromechanical composition. Finally, I reflect on the limitations of this project, and suggest directions for future research.

7.2 Summary

Algorithmic Transcription

In Chapter 4, addressing R1 on computationally extracting speech features and their implementation within a DIY practice involved a process of musical transformation. I used a small curated dataset, and a large corpus of emotional speech, with the differences between the speech source and mechanical realisation as a way to generate new musical material.

The choice of features—text-segmentation, spectral centroids, and Mel-spectrograms—could be seen to determine the output of the *Motorcase* devices. More specifically, the features and methods dictated which elements of the signal were preserved, and which were discarded. In practice, this data was mediated through motors and objects, a relationship which I described through a mosaic analogy: the speech source providing an overarching form, with the mechanical and acoustic qualities of the devices providing the surface detail. In this context, the controlled unpredictability of the mechanisms was essential to the work's DIY aesthetic.

I found that sustaining musical interest in performance required extending these transcriptions through performative controls, and created a tension between playback and intervention. Here, I used the original speech as a hidden driver for the behaviour of the system. With the aim of clarifying the link between speech and realisation, in each fixed composition I placed the devices in a small-ensemble setting, with *Years a have to know or something* using archival video and *Machine Translation* using RAVE-generated audio as a timbral enhancement. While my experience of playing these devices suggested they were best suited to improvisation, the methods provided musical material for both works that I would not have arrived at otherwise.

Audio Models

In Chapter 5, the integration of NAS as speech-trained RAVE models introduced a generative method for producing speech-derived material. This approach detached prosody and timbre

from human identity, treating speech as a decontextualised sonic material which I generated in bulk. To amplify the technological uncanny inherent in these models, I used poorly trained models, feedback loops and timbral transfer to emphasise artifacts and impose vocal qualities onto non-human sources. Transcribing these outputs to motors could be seen as furthering this abstraction, in pairing under-trained models with the unstable mechanical noise of found objects.

This abundance of generated material prompted two approaches to organisation: in *Smalltalk* I explored a contour-based search to group micro-gestures by similarity, informed by the non-semantic yet language-like forms of asemic writing. In *Radio User Guide*, I derived the form from an LSTM-generated text. Each provided me with a compositional constraint for the creative work, in the first case defining musical structures based on repeating patterns, and in the second around the gradual emergence of semantic meaning.

Integrating the materials with the RAVE-generated audio highlighted a mismatch between the timbral richness of neural synthesis and the limited range of the motors. While in *Smalltalk* I used a gestural shadowing that placed the motors in direct comparison with the voice, with *Radio User Guide* I integrated the generated audio by aligning the audio and physical objects with the generated text. A limitation across these works was the inference time and processing requirements of RAVE models, which restricted my use to pre-generated audio rather than real-time synthesis in the embedded context of the works.

Text Models

In Chapter 6 I addressed the affordances of text models by mapping linguistic structures—dialogue acts, dependency relations, and word embeddings—to mechanical action. Here, I used the organisation of language as a driver for musical structure, where the creative act was in the design of the mapping strategy, and the construction of each device.

In *Porcelain Music Box* and *Seven Studies*, I used text and language models to generate patterns in real-time. I found that the main affordance was the generation of new sequences that adhered to an internal logic. In my own work, this allowed the creation of conversational and syntactic patterns based on the flow of dialogue acts and dependency relations. In *Word games* I developed a method for translating semantic relationships into sound, by using the vector dimensions of a word to interpolate between musical parameter curves.

Limitations of this approach included the inevitable loss of data through translation to the DIY devices, particularly in the dimensionality reduction of word embeddings for *Word games*. In *Seven Studies*, my solution for the inference time of the embedded small language model was to loop or pause the output. Despite this, I would argue that in both *Porcelain Music Box* and *Seven Studies*, text models served as effective generators for algorithmic composition, providing structures derived from aspects of human communication. In this implementation, their value in my artistic practice was in generating material rather than as autonomous agents.

7.3 Practice Overview

To place each work within the context of the field reviews, the categorisation of sound sources and computational methods mirrors the *Discussion* from 2.5, followed by an overview of devices.

Computational Methods

Table 7.1 shows where each creative project employs the computational methods identified in Chapter 2. Several electronic manipulation categories were applied in an electromechanical context, such as the speed adjustment of *Motorcase I* or the looping of *Motorcase II*. Those not covered in this research include other forms of electronic manipulation—delays, reverb, granulation, pitch shifting—and the LPC method of speech synthesis, with *Radio User Guide* and *Smalltalk* using RAVE models instead.

Speech Sources

Table 7.2 shows the speech sources for each work, relative to those identified in Chapter 2. Speech corpus were used for both RAVE model training, and the navigation and playback of *Motorcase II*. The text-based-mappings are limited to the explorations of linguistic structure in Chapter 6. Sources not covered include audience and performer speech, field recordings and public figures, spoken literature and poetry. Part of the reason is a focus on the abstracted qualities of speech, which erased speaker identity and linguistic meaning, but these omissions give a clear indication for future work.

Devices and Interfaces

Chapter 2 covered instrumentations from orchestra to tape works. Every creative project in this research fits into the *Mechanical Devices* category, and so Table 7.3 instead summarises each device’s hardware, software, materials, and interface. The hardware stays relatively consistent and low-level (DC motors, contact mics, Arduino/Raspberry Pi), but the materials and interfaces vary across the works, from modular objects to found radio parts, and pre-composed playback to live-improvisation.

Table 7.1: Computational methods and works

Method	Subcategory	Work
Sample playback	<i>Fixed tape tracks</i>	Machine Translation, Smalltalk
	<i>Live triggering</i>	Years A Have to, Radio User Guide
Electronic manipulation	<i>Filtering</i>	Radio User Guide
	<i>Delays, reverb</i>	-
	<i>Loops and repetition</i>	Motorcase I, Motorcase II, Radio User Guide, Smalltalk, Word games
	<i>Granulation</i>	-
	<i>Interpolation</i>	Word games
	<i>Combination</i>	Motorcase I
	<i>Altering speed/direction</i>	Motorcase I
	<i>Pitch shifting</i>	-
Mapping	<i>Instrumental resynthesis</i>	Motorcase I, Motorcase II
	<i>Text-based or visual-</i>	Porcelain Music Box,
	<i>based mappings</i>	Seven Studies, Word games
Speech synthesis	<i>Linear Predictive Coding</i>	-
	<i>Machine learning</i>	Radio User Guide, Smalltalk

Table 7.2: Works by speech source.

Sound source	Work
Vocal formants	-
Casual conversation and interviews	Motorcase I, Motorcase II
Field recordings	-
Public figures	-
Spoken literature or poetry	-
Archival	Motorcase I
Performer's speech	-
Audience's speech	-
Affected speech	Motorcase II
Speech corpus	Motorcase II, Radio User Guide, Smalltalk
Text	Seven Studies, Word games, Radio User Guide, Porcelain Music Box
Grammar structures	Seven Studies, Word games, Porcelain Music Box
Synthesised speech	Radio User Guide, Smalltalk

Table 7.3: Summary of Devices.

	Hardware	Software	Objects	Interface
Motorcase I	8 DC motors, 2 contact mics, Arduino	MAX-MSP / Arduino	Mixed	Improvisation: playback / manipulation
Motorcase II	32 servo motors, 3 contact mics, Eurorack	Python / CircuitPy	Metal	Improvisation: navigation / manipulation
Smalltalk	2 DC motors, LEDs, Arduino, Rpi, Speaker	Python / Arduino	Metal	Playback (pre-composed)
Radio User Guide	3 DC motors, Rpi, Context mic, speaker	Python / Arduino	Radio parts	Live triggering (pre-composed)
Porcelain Music Box	12 DC motors, Rpi, Arduino	Python / Arduino	Porcelain	Installation: start / stop
Seven Studies	16 DC motors, Rpi, Arduino, LEDs, 2 contact mics, Eurorack	Python / Arduino / Puredata	Modular	Improvisation: playback / manipulation
Word games	9 DC motors, Rpi, Arduino, Camera, LEDs, 2 contact mics, Eurorack	Python / Arduino	Modular	Playback (pre-composed)

7.4 Themes

There were several shared themes across the creative works of the submitted portfolio, which I identified by reflecting on the project as a whole. These include approximation, material constraints, interaction, and transparency.

7.4.1 Approximation

Chapter 2 presented a field review of computational methods in speech-derived music. A recurring theme identified across these works was the deliberate obscuring of semantic meaning (2.5.3).

In my own practice, each work involved an approximation of speech, rather than a reproduction. This involved emphasising specific features, and discarding others. In *Motorcase I* and *II*, the method involved stripping away linguistic and prosodic detail to isolate particular sonic features: rhythmic, pitch-related, and amplitude-related. Each then involved further approximation, through the rhythmic text segmentation of *I*, or the spectral reduction to 32-bands in *II*.

In the RAVE-based works, the source was already a reconstruction created by a RAVE model. I chose to focus on emphasising the artifacts and failures of the model, making the approximation obvious. In my works exploring linguistic structure, the form was provided by the source, and my focus shifted to the material which was organised by this form; algorithmically generated or stochastically chosen from a large dataset. Here, it seemed that a speech-approximation was a result of this combination.

In each case, I used approximation as a method for generating novel musical material. As with the earliest examples of speech synthesis mentioned in 1.3, my aesthetic interest is in the difference between the source and realisation—how some features of speech remain, and others emerge through algorithmic and material approximation.

7.4.2 Material Constraints

Chapter 2 highlighted the theme of reproducing speech in a new medium, present in several existing works (2.5.3). In these existing works—such as Hummel’s *Nicanor*¹ or Riches’s *Talking*

1. Thomas Hummel, *Nicanor*, 1996, accessed 25 February 2026, <https://www.thomashummel.net/en/music/nicanor/>.

*Machine*²—while they still deal with the obscuring of speech, it seemed that the success of a particular method was often measured by the intelligibility of the reproduction. While I deliberately did not make intelligibility the aim in the works in the submitted portfolio, I did intend certain features in the original speech to survive each mapping process—such as the turn-taking qualities of *Porcelain Music Box*, or the semantic clustering of *Word games*.

Chapter 3 pointed to material constraints as a key aspect of existing electromechanical sound art (3.5.2), where the physical qualities of each device were both an affordance and limitation. Zareei's Audiovisual Materialism³ points to the inseparability of the audible, visual and material aspects of a work. This was most obvious in *Radio User Guide*, where the material qualities of the device were connected to the algorithmic and audio aspects of the work. In others, my impression was that limiting the objects to a single kind gave the works a coherence in a self-referential sense—as in the porcelain fragments of *Porcelain Music Box*. However I found that the most flexible and open-ended improvisatory practice was the result of the modular systems of both *Seven Studies* and *Word games*, which allowed me to explore a wider range of materials.

In all the works I use motors of some kind, from servo motors to DC motors, including the spindle variety salvaged from CD-players. While devices in Chapter 5 and 6 focus on rotational movement, devices in Chapter 4 turn rotational movement into linear movement and more unpredictable patterns, such as the bells of *Motorcase I*. This is partly because the devices in Chapter 4 focus on approximation as a generative method, whereas Chapter 6 is largely about the realisation of linguistic patterns, and I felt that a higher level of consistency was needed in this realisation. Other electromechanical components, such as solenoids, were not explored in this practice research and may form the basis for future systems.

Particular materials reoccur throughout the devices. This is partly a consequence of practical constraints: an actuator usually needs some flexibility if striking an object, and so I often use soft materials for this purpose including rubber bands, brush, thin wire, and fabric. Other objects appear in more than one device because I find their tactile and percussive sound appealing. This includes objects such as metal springs, copper sheets, paper and card. These are often not directly related to the underlying speech-derived process, and instead function as a realisation mechanism for the underlying system.

2. Martin Riches, *Talking Machine*, 1989, accessed 25 February 2026, <https://martinriches.de/>.

3. Mo H Zareei, 'Audiovisual materialism', *Organised Sound* 25, no. 3 (2020): 362–371, <https://doi.org/10.1017/s1355771820000321>.

The generative system and realisation mechanism here were often less tightly bound than with existing or more refined automatic instruments (3.5.2). For example, the modular objects of *Word games* function in realising the generated data, but did not affect it. Likewise, the metal objects of *Motorcase II* were driven by a speech-transcription, but were realised through such unstable object-actuators that the result could become decoupled from the transcription. For me, this allowed for the creation of new and unexpected musical material, but it also marked a point where material qualities could begin to override the features of the source. While this might seem like it defeats the purpose of the speech source for the system, I regard this process of decoupling interesting in itself, and my aesthetic choices throughout the project involved navigating this relationship between source, mapping, and realisation.

7.4.3 Interaction

In Chapter 2 I identified human-machine interaction as a recurring theme in existing speech-derived composition (2.5.3), and in Chapter 3 I developed this idea by categorising ML-based electromechanical works according to their mode of interaction. In the creative projects of the submitted portfolio, these modes ranged from the functional installation interface of *Porcelain Music Box* to the live shaping of improvisational output in *Seven Studies*.

In my improvisatory works, I developed interface controls through direct experimentation to establish an intuitive familiarity with the machine's behaviour. During performance, I found that interaction functioned as a reactive process; because of the stochastic qualities of the devices, I was often responding to unpredictable mechanical outputs.

However, there is a disconnect between my experience and the machine's logic. While the devices responded to interface data, they did not respond to their own output. The performer acts as the only link between the sound produced and the parameters being adjusted. While this process might be experienced as an act of co-creation, the system itself is one-directional. To me, this is an interesting result of the interactive systems I developed throughout the project, and a clear area for future work—where the devices could take audio or visual input from the output of their own process, to create feedback loops within the system.

7.4.4 Transparency

A recurring question in the submitted portfolio works was whether a source was audible in the realisation. Reductive and layered mapping strategies can create a situation where a speech source might become indistinguishable from abstract noise or purely algorithmic systems, and it is a valid critique that the outcome in works like *Porcelain Music Box* could be achieved through standard algorithmic procedures with similar aesthetic results. The systems I have explored here, from speech feature extraction to generating semantic labels, I find interesting from a compositional perspective. Making machines around these linguistic systems is a form of intellectual and artistic enjoyment, a kind of play. Here, the results of the process—the compositions, the improvisations, the physical devices—and how they are perceived from an aesthetic perspective is important, but as discussed in 6.5.4 it is not particularly important to me whether the system that drives them is perceived; whether the original sources of speech or language are visible, or not. I see the speech-derived process as a reductive translation, where audio signals or grammatical rules are distilled into mechanical behaviour. Within this process, I make creative decisions responding and attending to the structure of the data, and the emerging aesthetic qualities of the mechanical output. I find this to be a rewarding compositional process, with value in the moments of surprise, at the effect of a particular computational or material choice.

So what is the value of this work for other practitioners? A different artist might want, say, an audible sonification of conversational patterns. The practical questions they ask themselves might be much the same as those I have discussed—which dataset could be used, how could it be filtered, what model would be appropriate to generate these structures, what logic might dictate how these structures are interpreted, and so on. I hope that the documentation of each strategy will be of use in this respect, and each of my own implementations points to possible alternative uses of the same source.

7.5 Contributions

Compositional Strategies

The primary contributions are compositional strategies for bridging computationally-extracted speech features with physical sound production across three strands of practice. In Chapter 4 I

provide a way to use the divergence between speech-to-MIDI transcription and mechanical realisation as a generative tool, and a functional system for navigating a large speech dataset using emotional features. In Chapter 5 I introduce a strategy for selecting material from generated audio, and structuring a composition around generated text. Chapter 6 presents approaches to organising musical material around generated linguistic structures, and a musical interpretation of word embeddings.

Creative Work

This research includes a portfolio of seven original works and their corresponding devices. These works represent a practical application of the compositional strategies, ranging from fixed compositions to improvisatory performances using modular electromechanical systems.

Resources

This commentary provides two field reviews: a survey of computational methods in speech-derived music and an analysis of mapping strategies in mechanical sound art. The accompanying *Code* and *Models* folders contain the scripts, and trained models developed throughout the project. These are also made available online.^{4 5}

7.6 Limitations

The practice overview (7.3) mentioned areas not covered in this research, including several forms of electronic manipulation, and certain forms of speech synthesis. However the most significant omission was in live speech sources, including performer's and audience's speech.

Interaction with the devices remained one-way in the computational sense, with the performer controlling the machine output but the machines lacking audio or visual feedback loops that could develop the device's generative process. Similarly, the aim of portability and the resulting embedded approach to computation introduced some technical limitations: pre-generated RAVE audio rather than real-time, small text and language models rather than LLMs. Advances in microcontroller processing power, model distillation, and cloud-based inference will likely make these limitations easier to manage in future.

4. https://github.com/fifoh/-Speech_Derived_Music_Code/tree/main

5. <https://huggingface.co/fifoh/speech-derived-music-models/tree/main>

7.7 Future Work

These limitations suggest several areas for future work: the use of live speech sources, and the development of interactive systems that move beyond one-way control. This would involve routing audio from the device back into the system to modify the generative process in real-time, allowing the machine to respond its own output, and potentially extending to machine-machine interaction as a compositional strategy.

A further direction concerns the tension between improvisation and composition. The project highlighted the productive instability of speech-driven systems, but also the difficulty of translating such systems into reproducible works. I intend to explore the integration of DIY devices in musical ensemble contexts, particularly in relation to the modular approach to found objects. Within this, I see play and improvisation as central to developing these ideas.

Finally, I see the principles discussed here, of conversational flow as form, linguistic abstraction as structural logic, and approximation as creative method, extending beyond speech into other forms of algorithmic composition.

Appendices

Appendix A

Devices

A.1 Motorcase I

Dimensions

25 cm × 30 cm × 60 cm

Materials

Hard shell case, ceramic, glass, stone, metal grill, spoon, beads, metal plate, chain, brass bells.

Hardware

8 DC motors, 2 contact microphones, Arduino Uno.



Figure A.1: Motorcase I

A.2 Motorcase II

Dimensions

20 cm × 30 cm × 50 cm

Materials

Hard shell case, springs, heatsinks, metal plates, toy piano bars, rubber bands, wooden beads.

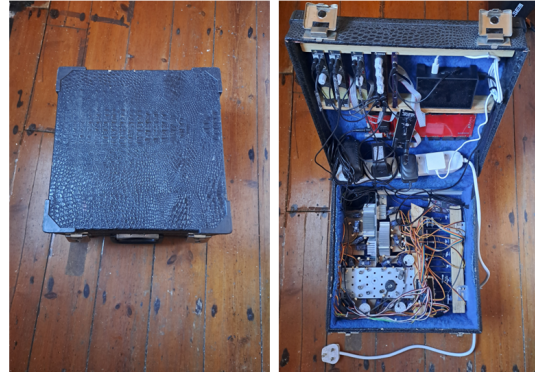


Figure A.2: Motorcase II

Hardware

32 servo motors, 3 contact microphones, Grand Central microcontroller, Eurorack.

A.3 Smalltalk

Dimensions

Three devices, each 25 cm × 20 cm × 20 cm

Materials

Wooden box, metal plates, springs.

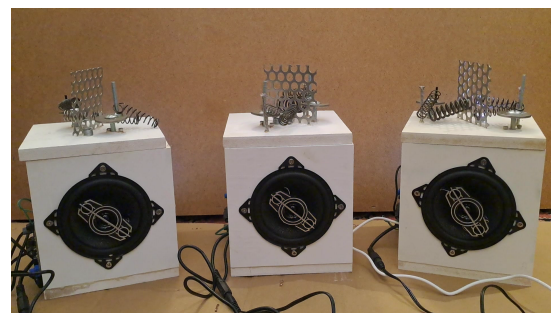


Figure A.3: Smalltalk

Hardware

2 DC motors, 2 LEDs, speaker, amplifier, Raspberry Pi, Arduino Uno.

A.4 Radio User Guide

Dimensions

Three devices, each 40 cm × 20 cm × 10 cm

Materials

Perspex panel, mixed radio parts.

Hardware

3 DC motors, 2 speakers, 2 amplifiers, contact microphone, Raspberry Pi.

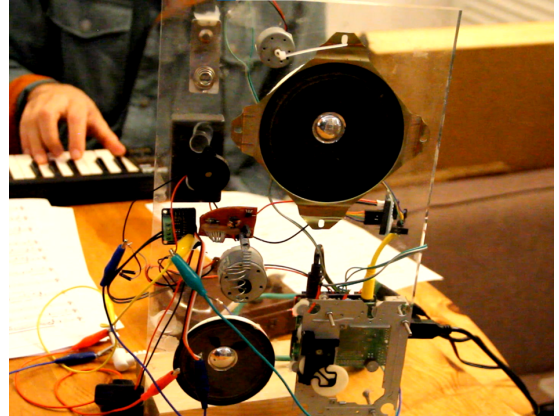


Figure A.4: Radio User Guide

A.5 Porcelain Music Box

Dimensions

45 cm × 30 cm × 25 cm

Materials

Wooden box, porcelain, cable ties, guitar strings.

Hardware

12 DC motors, Raspberry Pi, Arduino Uno.

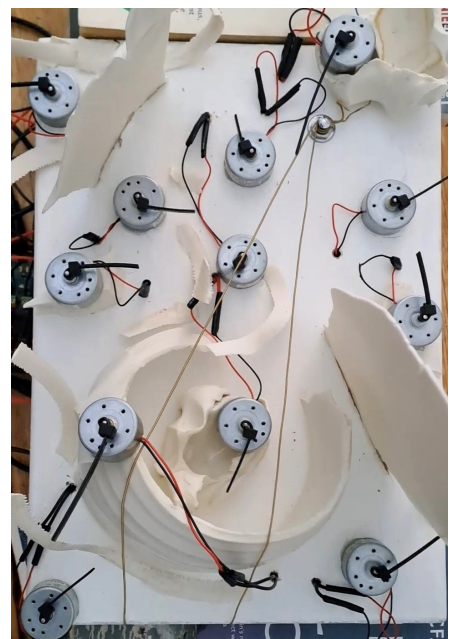


Figure A.5: Porcelain Music Box

A.6 Seven Studies

Dimensions

55 cm × 40 cm × 25 cm

Materials

Perspex panel, mixed objects.

Hardware

16 DC motors, 2 contact microphones, Raspberry Pi, Eurorack.

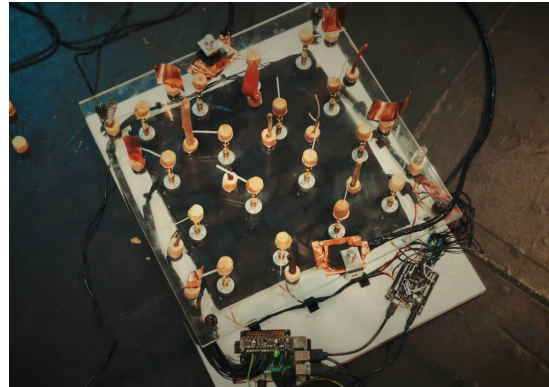


Figure A.6: Seven Studies

A.7 Word Games

Dimensions

35 cm × 55 cm × 25 cm

Materials

Wooden box, mixed objects, clips, 2 contact microphones, Eurorack.

Hardware

9 DC motors, Raspberry Pi, Arduino Uno, camera.



Figure A.7: Word Games

References

- Abadi, Martín, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard et al. ‘TensorFlow: A system for Large-Scale machine learning’. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, 265–283. 2016. <https://doi.org/10.48550/arXiv.1605.08695>.
- Ablinger, Peter. *Quadraturen*, 1995 -. Accessed 25 February 2026. <https://ablinger.mur.at/docu11.html>.
- Ablinger, Peter. *Quadraturen II*, 2002. Accessed 25 February 2026. <https://ablinger.mur.at/docu11.html#qu2>.
- Ablinger, Peter. *Quadraturen III*, 2004. Accessed 25 February 2026. <https://ablinger.mur.at/docu11.html#qu3>.
- Ablinger, Peter. *Voices and Piano*, 1998-2003. Accessed 25 February 2026. https://ablinger.mur.at/docs/v+p_ginsburgh.pdf.
- Adámek, Ondrej. *Ça tourne ça bloque*, 2007. Accessed 25 February 2026. <https://ondrejadamek.com/listen/ca-tourne-ca-bloque-10-musiciansssamplervideo/>.
- Akten, Memo and Jennifer Walshe. *ULTRACHUNK*, 2018. Accessed 25 February 2026. <https://memo.tv/works/ultrachunk/>.
- Allal, Loubna Ben, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav et al. ‘SmolLM2: When Smol Goes Big—Data-Centric Training of a Small Language Model’. *arXiv preprint arXiv:2502.02737*, 2025. <https://doi.org/10.48550/arXiv.2502.02737>.
- Almeida, Felipe and Geraldo Xexéo. ‘Word embeddings: A survey’. *arXiv preprint arXiv:1901.09069*, 2019. <https://doi.org/10.48550/arXiv.1901.09069>.
- Amari, Shun-ichi. ‘Dynamics of pattern formation in lateral-inhibition type neural fields’. *Biological cybernetics* 27, no. 2 (1977): 77–87. <https://doi.org/10.1007/BF00337259>.
- Anhalt, Istvan. ‘Alternative Voices: Essays on contemporary vocal and choral composition’. In *Alternative Voices*. University of Toronto Press, 1984. <https://doi.org/10.3138/9781442656314>.
- Armitage, Jack and Victor Shepardson. ‘Augmenting the expressivity of the Notochord generative midi model for Arca’s “The Light Comes in the Name of the Voice”’. In *Proceedings of the 5th Conference on AI Music Creativity*. 2024. Accessed 25 February 2026. <https://aimc2024.pubpub.org/pub/0lh6s86c>.

- Astrinaki, Maria, Nicolas d'Alessandro, Loic Reboursière, Alexis Moinet and Thierry Dutoit. 'MAGE 2.0: New Features and its Application in the Development of a Talking Guitar'. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, 547–550. 2013. <https://doi.org/https://doi.org/10.5281/zenodo.1178467>.
- Audry, Sofian, Victor Drouin-Trempe and Ola Siebert. 'The Strangest Music in the World: Self-Supervised Creativity and Nostalgia for the Future in Robotic Rock Band "The Three Sirens"'. In *Arts*, 12:2. 1. MDPI, 2022. <https://doi.org/10.3390/arts12010002>.
- Bacon, Benjamin. *Little Sound Machines*, 2018. Accessed 25 February 2026. <https://www.benjaminbacon.studio/little-sound-machines>.
- Bacon, Benjamin. *PROBE II, Subaudition*, 2021. Accessed 25 February 2026. <https://www.benjaminbacon.studio/probe-1>.
- Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed and Michael Auli. 'wav2vec 2.0: A framework for self-supervised learning of speech representations'. *Advances in neural information processing systems* 33 (2020): 12449–12460. <https://doi.org/10.48550/arXiv.2006.11477>.
- Bailes, Freya and Roger T Dean. 'When is noise speech? A survey in sonic ambiguity'. *Computer Music Journal* 33, no. 1 (2009): 57–67. <https://doi.org/10.1162/comj.2009.33.1.57>.
- Bailie, Joanna. *Artificial Environments no.8*, 2013. Accessed 25 February 2026. <https://joanna.bailie.com/music/solo-and-small-chamber/ae8/>.
- Barlow, Clarence. *Farting Quietly in Church*, 1992. Accessed 25 February 2026. <https://clarlow.org/compositions-by-year>.
- Barlow, Clarence. *fLvXvS*, 1990. Accessed 25 February 2026. <https://clarlow.org/compositions-by-year>.
- Barlow, Clarence. 'On Music Derived from Language'. In *Proceedings of the 3rd International Symposium on Systems Research in the Arts and Humanities, Baden-Baden, Germany*. 2009. Accessed 25 February 2026. <https://clarlow.org/wp-content/uploads/2025/06/On-Music-Derived-from-Language.pdf>.
- Barlow, Clarence. *Orchideæ Ordinariæ*, 1989. Accessed 25 February 2026. <https://www.soundohm.com/product/musica-algorithmica-2-cd>.
- Barlow, Clarence. *Progéthal Percussionian for the Advanced Beginners*, 2003. Accessed 25 February 2026. <https://clarlow.org/compositions-by-year/>.
- Bartley, W. *Concert Report: The Fine Art of Social Consciousness*, 2017. Accessed 25 February 2026. <https://thewholenote.com/index.php/newsroom/blog/concert-reviews/27382-concert-report-the-fine-art-of-social-consciousness-james-tenney-resistance>.
- Bartley, Wende. *Ellipsis*, 1988. Accessed 25 February 2026. <https://electrocd.com/en/oeuvre/14175/wende-bartley/ellipsis>.
- Becker, Jonas, Jan Philip Wahle, Bela Gipp and Terry Ruas. 'Text generation: A systematic literature review of tasks, evaluation, and challenges'. *arXiv preprint arXiv:2405.15604*, 2024. <https://doi.org/10.48550/arXiv.2405.15604>.

- Bittner, Rachel M., Juan José Bosch, David Rubinstein, Gabriel Meseguer-Brocal and Sebastian Ewert. 'A Lightweight Instrument-Agnostic Model for Polyphonic Note Transcription and Multipitch Estimation'. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Singapore, 2022. <https://doi.org/10.1109/ICASSP43922.2022.9746549>.
- Blablabor. *hirsch hirn hornisse*, 2009. Accessed 25 February 2026. <https://www.blablabor.ch/hirsch-hirn-hornisse>.
- Blonk, Jaap. *Traces of Speech*, 2014. Accessed 25 February 2026. <https://jaapblonk.bandcamp.com/album/traces-of-speech>.
- Bogaards, Niels and Axel Röbel. 'An interface for analysis-driven sound processing'. In *AES 119th Convention*, 1–1. 2005. Accessed 25 February 2026. <https://hal.science/hal-01161331v1>.
- Bolt, Barbara. 'The Magic is in Handling'. In *Practice as research: Approaches to creative arts enquiry*, edited by Estelle Barrett and Barbara Bolt, 27–34. Ib Tauris London, 2007. <https://doi.org/10.5040/9780755604104.ch-002>.
- Boselli, Roberto, Simone D'Amico and Navid Nobani. 'eXplainable AI for word embeddings: a survey'. *Cognitive Computation* 17, no. 1 (2025): 19. <https://doi.org/10.1007/s12559-024-10373-2>.
- Bosetti, Alessandro. *Alpine Flipbooks*, 2013. Accessed 25 February 2026. <https://melgun.net/travaux/alpine-flipbooks/>.
- Bosetti, Alessandro. *Mask Mirror*, 2008. Accessed 25 February 2026. <https://melgun.net/travaux/maskmirror/>.
- Bosetti, Alessandro. *The Notebooks*, 2014. Accessed 25 February 2026. <https://melgun.net/travaux/the-notebooks/>.
- Bossis, Bruno. 'Reflections on the analysis of artificial vocality: representations, tools and perspective'. *Organised Sound* 9, no. 1 (2004): 91–98. <https://doi.org/10.1017/S1355771804000123>.
- Bowers, John and Phil Archer. 'Not hyper, not meta, not cyber but infra-instruments'. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, 5–10. 2005. <https://doi.org/10.5281/zenodo.1176713>.
- Bresson, Jean, Carlos Agon and Gérard Assayag. 'OpenMusic: visual programming environment for music composition, analysis and research'. In *Proceedings of the 19th ACM international conference on Multimedia*, 743–746. 2011. <https://doi.org/10.1145/2072298.2072434>.
- Bretan, Mason and Gil Weinberg. 'A survey of robotic musicianship'. *Communications of the ACM* 59, no. 5 (2016): 100–109. <https://doi.org/10.1145/2818994>.
- Briot, Jean-Pierre. 'From artificial neural networks to deep learning for music generation: history, concepts and trends'. *Neural Computing and Applications* 33, no. 1 (2021): 39–65. <https://doi.org/10.1007/s00521-020-05399-0>.
- Brown, Andrew and John Ferguson. 'DIY musical instruments: From Handmade Electronic Circuits to Microcontrollers and Digital Fabrication'. *Journal of Ubiquitous Music*, 2024. Accessed 25 February 2026. <https://hdl.handle.net/10072/434542>.

- Bugallo, Helena. ‘Selected “Studies” for player piano by Conlon Nancarrow: Sources, working methods, and compositional strategies’. PhD diss., State University of New York at Buffalo, 2004.
- Bunt, Harry. ‘Dimensions in Dialogue Act Annotation.’ In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, 919–924. European Language Resources Association, 2006. Accessed 25 February 2026. http://www.lrec-conf.org/proceedings/lrec2006/pdf/428_pdf.pdf.
- Caillon, Antoine and Philippe Esling. ‘RAVE: A variational autoencoder for fast and high-quality neural audio synthesis’. *CoRR*, 2021. <https://doi.org/10.48550/arXiv.2111.05011>.
- Caillon, Antoine and Philippe Esling. *RAVE: Official implementation of the RAVE model*, 2021. Accessed 25 February 2026. <https://github.com/acids-ircam/RAVE>.
- Canuto, Jesus. *Talking Motors*, 2023. Accessed 25 February 2026. <https://www.jesuscanuto.com/talking-motors>.
- Cardilli, Lorenzo, Stefano Lombardi Vallauri et al. *L’arte orale. Poesia, musica, performance*. Accademia University Press, 2020. <https://doi.org/10.4000/books.aaccademia.7908>.
- Carpentier, Grégoire, Gérard Assayag and Emmanuel Saint-James. ‘Solving the musical orchestration problem using multiobjective constrained optimization with a genetic local search approach’. *Journal of Heuristics* 16, no. 5 (2010): 681–714. <https://doi.org/10.1007/s10732-009-9113-7>.
- Chadburn, Leo. ‘Captions, characters, self-portraits: compositional approaches to the disembodied speaking voice and the voice-text-music relationship’. PhD diss., City, University of London, 2023.
- Chao, Noelle. ‘Listening to the Voice on the Page: Joshua Steele and Technologies of Recording’. *The Eighteenth Century* 54, no. 2 (2013): 245–261. <https://doi.org/10.1353/ecy.2013.0012>.
- Chau, Christina. ‘Movement and Time in the Nexus between Technological Modes with Jean Tinguely’s Kineticism’. In *Arts*, 3:394–406. 4. MDPI, 2014. <https://doi.org/10.3390/arts3040394>.
- Chen, Andrew. ‘Voice Without Speaker: human speech synthesis in acoustic instrumental contexts’. *Tempo* 77, no. 305 (2023): 60–71. <https://doi.org/10.1017/s0040298223000098>.
- Ciardi, Fabio Cifariello. ‘Strategies and tools for the sonification of prosodic data: A composer’s perspective’. In *Proceedings of the 26th International Conference on Auditory Display (ICAD 2021)*, 247–252. 2021. <https://doi.org/10.21785/icad2021.041>.
- Code, David Loberg. ‘Observations in the Art of Speech: Paul Lansky’s Six Fantasies’. *Perspectives of New Music*, 1990, 144–169. <https://doi.org/10.2307/833348>.
- Couchot, Edmond. ‘Automatism, Autonomy and Aesthetics in the Performing Arts’. *Leonardo* 52, no. 3 (2019): 240–246. https://doi.org/10.1162/leon_a_01752.
- Coutinho, Eduardo and Nicola Dibben. ‘Psychoacoustic cues to emotion in speech prosody and music’. *Cognition & emotion* 27, no. 4 (2013): 658–684. <https://doi.org/10.1080/02699931.2012.732559>.
- Cummins, Fred, Marco Grimaldi, Thomas Leonard and Juraj Simko. ‘The chains speech corpus: Characterizing individual speakers’. In *Proc of SPECOM*, 1–6. 2006. <https://doi.org/10.35111/cgbv-ke56>.

- De Marneffe, Marie-Catherine, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre and Christopher D Manning. ‘Universal Stanford dependencies: A cross-linguistic typology.’ In *LREC*, 14:4585–4592. 2014. Accessed 25 February 2026. http://www.lrec-conf.org/proceedings/lrec2014/pdf/1062_Paper.pdf.
- Dean, Roger. ‘NoiseSpeech, a Noise of Living Bodies: Towards Attali’s’ Composition”. *Journal of New Media and Culture* 3, no. 2 (2005): 1–12. Accessed 25 February 2026. <https://ibiblio.org/nmediac/winter2004/NoiseSpc.htm>.
- Debicki, G. and D. Gardener. *Empty Vessels*, 2019. Accessed 25 February 2026. <https://ars.electronica.art/newdigitaldeal/en/empty-vessels/>.
- Defferrard, Michaël, Kirell Benzi, Pierre Vandergheynst and Xavier Bresson. ‘FMA: A dataset for music analysis’. *arXiv preprint arXiv:1612.01840*, 2016. <https://doi.org/10.48550/arXiv.1612.01840>.
- DeMarinis, Paul. *Alien Voices*, 1988. Accessed 25 February 2026. <https://pauldemarinis.org/AlienVoices.html>.
- DeMarinis, Paul. *Music as a Second Language*, 1991. Accessed 25 February 2026. <https://soundohm.com/product/music-as-a-second-language-2>.
- Donahue, Chris, Ian Simon and Sander Dieleman. ‘Piano genie’. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 160–164. 2019. <https://doi.org/10.1145/3301275.3302288>.
- Doornbusch, Paul. ‘Composers’ views on mapping in algorithmic composition’. *Organised sound* 7, no. 2 (2002): 145–156. <https://doi.org/10.1017/s1355771802002066>.
- Dunham, Paul. ‘Click:: TWEET. A Telegraphic Sound Installation’. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 2021. <https://doi.org/10.21428/92fb44.48c39866>.
- Dyer, Mark, Zubin Kanga and Jennifer Walshe. ‘In Conversation with Jennifer Walshe: Performing with Intelligent Machines’. *Contemporary Music Review* 42, no. 3 (2023): 391–399. <https://doi.org/10.1080/07494467.2023.2276563>.
- Eacott, John. ‘Flood Tide: sonification as musical performance—an audience perspective’. *AI & society* 27 (2012): 189–195. <https://doi.org/10.1007/s00146-011-0338-2>.
- Emard, Justine. *Supraorganism*, 2020. Accessed 25 February 2026. <https://justineemard.com/supraorganism/>.
- Evans, Nicholas, Behzad Haki and Sergi Jordà. ‘Repurposing a Rhythm Accompaniment System for Pipe Organ Performance’. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, 116–120. 2025. <https://doi.org/10.5281/zenodo.15698807>.
- Faubel, C. and W. Spahn. *Symbolic Grounding*, 2018. Accessed 25 February 2026. <https://christian.faubel.derstrudel.org/symbolic-grounding/>.
- Faubel, C. and W. Spahn. *Whispering*, 2016–2022. Accessed 25 February 2026. <https://www.alimiharbi.com/work/whisper/>.
- Faubel, Christian. ‘Emergent Polyrhythmic Patterns with a Neuromorph Electronic Network’. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 2021. <https://doi.org/10.21428/92fb44.e66a8542>.

- Ferreira, Pedro, Ricardo Limongi and Luiz Paulo Fávero. 'Generating music with data: application of deep learning models for symbolic music composition'. *Applied Sciences* 13, no. 7 (2023): 4543. <https://doi.org/10.3390/app13074543>.
- Fiebrink, Rebecca and Perry R Cook. 'The Wekinator: a system for real-time, interactive machine learning in music'. In *Proceedings of the eleventh international society for music information retrieval conference*, 3:2–1. International Society for Music Information Retrieval, 2010. Accessed 25 February 2026. <https://ismir2010.ismir.net/proceedings/late-breaking-demo-13.pdf>.
- Fiebrink, Rebecca and Laetitia Sonami. 'Reflections on eight years of instrument creation with machine learning'. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, 237–242. 2020. Accessed 25 February 2026. <https://research.gold.ac.uk/id/eprint/28405>.
- Flø, Asbjørn Blokkum. 'Materiality in sound art'. *Organised Sound* 23, no. 3 (2018): 225–234. <https://doi.org/10.1017/s1355771818000134>.
- Ford, Corey, Ashley Noel-Hirst, Sara Cardinale, Jackson Loth, Pedro Sarmiento, Elizabeth Wilson, Lewis Wolstanholme, Kyle Worrall and Nick Bryan-Kinns. 'Reflection Across AI-based Music Composition'. In *Proceedings of the 16th Conference on Creativity & Cognition*, 398–412. 2024. <https://doi.org/10.1145/3635636.3656185>.
- Formo, Daniel Buner. 'The Orchestra of Speech—a speech-based instrument system'. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 2018. <https://doi.org/10.5281/zenodo.1302607>.
- Franklin, Austin. 'The Robo-Cajon: An Example of Live Performance with Musical Robotics'. In *International Computer Music Conference, ICMC Proceedings*, 270–273. International Computer Music Association, 2024. Accessed 25 February 2026. <https://urn.kb.se/resolve?urn=urn:nbn:se:mdh:diva-70059>.
- Fu, Tairan, Raquel Ferrando, Javier Conde, Carlos Arriaga and Pedro Reviriego. 'Why Do Large Language Models (LLMs) Struggle to Count Letters?' *arXiv preprint arXiv:2412.18626*, 2024. <https://doi.org/10.48550/arXiv.2412.18626>.
- Furrer, Beat. *Wüstenbuch*, 2009. Accessed 25 February 2026. <https://www.kairos-music.com/cds/0013312kai>.
- Gadenstätter, Clemens. *Semantical Investigations II*, 2007. Accessed 25 February 2026. <https://www.kairos-music.com/cds/0015006kai>.
- Gál, Bernhard. *bestimmung new york*, 1999. Accessed 25 February 2026. <https://bernhardgal.com/bestimmungnewyork.html#info>.
- Gioti, Artemi-Maria. 'Imitation game: Real-time decision-making in an inter-active composition for human and robotic percussionist'. In *Proceedings of the International Computer Music Conference*. 2019. Accessed 25 February 2026. https://interagency.iem.at/pdfs/Gioti_ICMC2019.pdf.
- Gold, Bernard, Nelson Morgan and Daniel Ellis. *Speech and audio signal processing: processing and perception of speech and music*. Wiley, 2011.
- Graves, Alex. 'Generating sequences with recurrent neural networks'. *arXiv preprint arXiv:1308.0850*, 2013. <https://doi.org/10.48550/arXiv.1308.0850>.

- Gries, Stefan Th. and Andrea L. Berez. 'Linguistic Annotation in/for Corpus Linguistics'. In *Handbook of Linguistic Annotation*, edited by Nancy Ide and James Pustejovsky, 379–409. Springer, 2017. https://doi.org/10.1007/978-94-024-0881-2_15.
- Grill, T. *Mutual Understanding*, 2019. Accessed 25 February 2026. https://rottingsounds.org/threads/projects/mutual_understanding/.
- Grouse, Matthew. *Left Right, Left Right*, 2020. Accessed 25 February 2026. <https://www.matthewgrousemusic.com/left-right-left-right>.
- Haki, Behzad, Nicholas Evans and Sergi Jordà. 'GrooveTransformer: A Generative Drum Sequencer Eurorack Module'. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, 261–265. 2024. <https://doi.org/10.5281/zenodo.13904848>.
- Hallam, Elizabeth and Tim Ingold. *Creativity and cultural improvisation*. Routledge, 2021.
- Harriman, Jiffer, Matthew Bethancourt, Abhishek Narula, Michael Theodore and Mark Gross. 'SolidNoise: Making Musical Robots'. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2504–2510. 2016. <https://doi.org/10.1145/2851581.2892539>.
- Harvey, Jonathan. *Speakings*, 2007. Accessed 25 February 2026. <https://www.fabermusic.com/shop/speakings-p410799>.
- Hausmann, Raoul. *f m s b w t ö z ä u*, 1918. Accessed 25 February 2026. <https://www.centrepompidou.fr/fr/ressources/oeuvre/cozKBB>.
- Hecker, Florian. *Inspection II*, 2019. Accessed 25 February 2026. <https://lampo.org/archive/florian-hecker-2019/>.
- Hellstenius, Henrik. *Instrument of Speech*, 2016. Accessed 25 February 2026. https://hellstenius.no/wp-content/uploads/2018/09/PLING_An-expanded-palette-2.pdf.
- Helmholtz, M. 'XII. On vowel sounds'. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 19, no. 125 (1860): 81–88. <https://doi.org/10.1080/14786446008642829>.
- Hodkinson, Juliana. 'Resounding difficult histories'. In *Over and Over and Over Again: Reenactment Strategies in Contemporary Arts and Theory*, edited by Cristina Baldacci, Clio Nicastro and Arianna Sforzini, 69–80. ICI Berlin Press, 2022.
- Holbrow, Charles, Elena Jessop and Rébecca Kleinberger. 'Vocal vibrations: A Multisensory Experience of the Voice'. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, 431–434. 2014. <https://doi.org/10.5281/zenodo.1178800>.
- Holmes, Thom. *Electronic and experimental music: technology, music, and culture*. Routledge, 2012.
- Honnibal, Matthew, Ines Montani, Sofie Van Landeghem and Adriane Boyd. 'spaCy: Industrial-strength Natural Language Processing in Python', 2020. <https://doi.org/10.5281/zenodo.1212303>.
- Horio, Kanta. *Particle*, 2008. Accessed 25 February 2026. <https://horiokanta.com/particle/>.
- Horta Valenzuela, Moisés. *Rave Latent Diffusion*, 2023. Accessed 25 February 2026. <https://github.com/moiseshorta/RAVE-Latent-Diffusion>.

- Hsu, Aurie and Steven Kemper. ‘Enacting Sonic-Cyborg Performance through the Hybrid Body in Teka-Mori and Why Should Our Bodies End at the Skin?’ *Leonardo Music Journal* 29 (2019): 83–87. https://doi.org/10.1162/lmj_a_01069.
- Hu, Di and Katunobu Ito. ‘Two-stage instrument timbre transfer method using RAVE’. In *2024 International Symposium on Multimedia (ISM)*, 33–40. 2024. <https://doi.org/10.1109/ISM63611.2024.00011>.
- Huang, Ary-Yue. ‘Artificial Life, One Leg at a Time: The Aesthetics of Trial-and-Error in AI Training’. In *Proceedings of the 2025 Conference on Creativity and Cognition*, 818–821. 2025. <https://doi.org/10.1145/3698061.3726944>.
- Hummel, Thomas. *Nicanor*, 1996. Accessed 25 February 2026. <https://www.thomashummel.net/en/music/nicanor/>.
- Hymel, Shawn, Colby Banbury, Daniel Situnayake, Alex Ellum, Carl Ward, Mat Kelcey, Mathijs Baaijens, Mateusz Majchrzycki, Jenny Plunkett, David Tischler et al. ‘Edge impulse: An mlops platform for tiny machine learning’. *arXiv preprint arXiv:2212.03332*, 2022. <https://doi.org/10.48550/arXiv.2212.03332>.
- Jensenius, Alexander Refsum and Marcelo M Wanderley. ‘Musical gestures: Concepts and methods in research’. In *Musical gestures*, edited by Rolf Inge Godøy and Marc Leman, 24–47. Routledge, 2010.
- Ji, Shulei, Xinyu Yang and Jing Luo. ‘A survey on deep learning for symbolic music generation: Representations, algorithms, evaluations, and challenges’. *ACM Computing Surveys* 56, no. 1 (2023): 1–39. <https://doi.org/10.1145/3597493>.
- Jung, Matthias. ‘Intelligent Music Performance Systems: Towards a Design Framework’. *Studia Musicologica Norvegica* 49, no. 1 (2023): 28–44. <https://doi.org/10.18261/smn.49.1.3>.
- Jurafsky, Daniel and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*. 3rd. 2026. Accessed 25 February 2026. <https://web.stanford.edu/~jurafsky/slp3/>.
- Juslin, Patrik N and Petri Laukka. ‘Communication of emotions in vocal expression and music performance: Different channels, same code?’ *Psychological bulletin* 129, no. 5 (2003): 770.
- Kapur, Ajay, Michael Darling, Dimitri Diakopoulos, Jim W Murphy, Jordan Hochenbaum, Owen Vallis and Curtis Bahn. ‘The machine orchestra: An ensemble of human laptop performers and robotic musical instruments’. *Computer Music Journal* 35, no. 4 (2011): 49–63. https://doi.org/10.1162/comj_a_00090.
- Kirath, Şöl en, Hannah E Wolfe and Alex Bundy. ‘Cacophonic choir: an interactive art installation embodying the voices of sexual assault survivors’. In *ACM SIGGRAPH 2020 Art Gallery*, 446–450. Association for Computing Machinery, 2020. <https://doi.org/10.1145/3386567>.
- Kleinberger, Rébecca, Nikhil Singh, Xiao Xiao and Akito van Troyer. ‘Voice at NIME: a Taxonomy of New Interfaces for Vocal Musical Expression’. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 2022. <https://doi.org/10.21428/92fbeb44.4308fb94>.

- Ko, Jiin, Youngjun Choi and Jinjoon Lee. 'Harmonic Words, Narrative Chords: Textual Sonification Using Part of Speech'. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, 28–32. 2024. <https://doi.org/10.5281/zenodo.13904771>.
- Kopito, Gabriel, Jonathan Schwartz, Julien Amblard, Robert Filman and Landon Rabern. 'Ghost in the Piano – An Investigation into AI, Robotics and Musical Performance'. In *AIMC 2024*. 2024. <https://doi.org/10.5281/zenodo.15110248>.
- Kress, Gregg. *Brain Noise Machine*, 2010. Accessed 25 February 2026. <https://glkress.com/art-and-design/brain-art/brain-noise-machine/>.
- Lane, Cathy. 'Voices from the Past: Compositional approaches to using recorded speech'. *Organised Sound* 11, no. 1 (2006): 3–11. <https://doi.org/10.1017/s1355771806000021>.
- Lansky, Paul. *Word Color*, 1992. Accessed 25 February 2026. https://paul.mycpanel.princeton.edu/liner_notes/morethanidlechatter.html#.
- Lanza, Mauro. *The Kempelen Machine*, 2015. Accessed 25 February 2026. <https://brahms4.ircam.fr/fr/works/work/41088/>.
- Leitman, Sasha. 'Trimpin: an interview'. *Computer Music Journal* 35, no. 4 (2011): 12–27. https://doi.org/10.1162/comj_a_00088.
- Lewis, Luke. *The Echoes Return Slow*, 2021. Accessed 25 February 2026. <https://luke-lewis.com/music/theechoesreturnslow.html>.
- Li, Junyi, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie and Ji-Rong Wen. 'Pre-trained language models for text generation: A survey'. *ACM Computing Surveys* 56, no. 9 (2024): 1–39. <https://doi.org/10.1145/3649449>.
- Livingstone, Steven R and Frank A Russo. 'The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English'. *PloS one* 13, no. 5 (2018): e0196391. <https://doi.org/10.1371/journal.pone.0196391>.
- Lotfian, Reza and Carlos Busso. 'Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings'. *IEEE Transactions on Affective Computing* 10, no. 4 (2017): 471–483. <https://doi.org/10.1109/TAFFC.2017.2736999>.
- Mâche, François-Bernard. *Korwar*, 1972. Accessed 25 February 2026. <https://brahms.ircam.fr/work/korwar>.
- Macionis, McLean J and Ajay Kapur. 'Where is the quiet: Immersive experience design using the brain, mechatronics, and machine learning'. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, 335–338. 2019. <https://doi.org/10.5281/zenodo.3672984>.
- Malherbe, Claudy. *Locus*, 1997. Accessed 25 February 2026. <https://brahms.ircam.fr/en/works/work/10464/>.
- Malherbe, Claudy. 'Locus: Rien n'aura eu lieu que le lieu'. In *The OM Composer's Book 2*, edited by Jean Bresson, Carlos Agon and Gérard Assayag. Delatour/Ircam, 2008.
- Maltman, Jonathan and Lindsay Vickery. 'The Foundations of Speech-Melody Composition And The Dawn Of The Digital Music Era'. In *Proceedings of the Australasian Computer Music Conference*, 30–37. 2018.

- Manovich, Lev. *Data Visualization as New Abstraction and Anti-Sublime*, 2002. Accessed 25 February 2026. https://manovich.net/content/04-projects/041-data-visualisation-as-new-abstraction-and-anti-sublime/37_article_2002.pdf.
- Marcondes, Francisco S, Adelino Gala, Renata Magalhães, Fernando Perez de Britto, Dalila Durães and Paulo Novais. ‘Using ollama’. In *Natural Language Analytics with Generative Large-Language Models: A Practical Approach with Ollama and Open-Source LLMs*, 23–35. Springer, 2025. https://doi.org/10.1007/978-3-031-76631-2_3.
- Marcus, Mitch, Beatrice Santorini and Mary Ann Marcinkiewicz. ‘Building a large annotated corpus of English: The Penn Treebank’. *Computational linguistics* 19, no. 2 (1993): 313–330. <https://doi.org/10.21236/ada273556>.
- McGuire, Riley. ‘Writing Novels, Simulating Voices: Euphonia, Trilby, and the Technological Sounding of Identity’. *Victorian Literature and Culture* 49, no. 2 (2021): 325–360. <https://doi.org/10.1017/s1060150319000251>.
- McInnes, Leland, John Healy and James Melville. ‘Umap: Uniform manifold approximation and projection for dimension reduction’. *arXiv preprint arXiv:1802.03426*, 2018. <https://doi.org/10.48550/arXiv.1802.03426>.
- McPherson, Andrew. ‘The magnetic resonator piano: Electronic augmentation of an acoustic grand piano’. *Journal of New Music Research* 39, no. 3 (2010): 189–202. <https://doi.org/10.1080/09298211003695587>.
- Miller, Cassandra. ‘Transformative Mimicry: Composition as Embodied Practice in Recent Works’. PhD diss., University of Huddersfield, 2018.
- Mitra, Rohan and Imran Zualkernan. ‘Music generation using deep learning and generative AI: a systematic review’. *IEEE Access*, 2025. <https://doi.org/10.1109/access.2025.3531798>.
- Muenz, Harald. *deChiffraGE*, 2004. Accessed 25 February 2026. <https://www.musicwedliketoh ear.com/pdf/29032006.pdf>.
- Müller, Meinard. ‘Dynamic Time Warping’. In *Information Retrieval for Music and Motion*, 69–84. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. https://doi.org/10.1007/978-3-540-74048-3_4.
- Nikrang, A. *Absolute Hallucination*, 2024. Accessed 25 February 2026. [https://ars.electronica.art/futurelab/en/projects-absolute-hallucination/..](https://ars.electronica.art/futurelab/en/projects-absolute-hallucination/)
- Nikrang, Ali and Susanne Kiesenhofer. ‘Human–AI Co-Creation in Contemporary Composition: Interaction and Artistic Strategies with Ricercar’. In *Proceedings of the Conference on Animation and Interactive Art*, 65–73. 2025. <https://doi.org/10.1145/3749893.3749961>.
- Ning, Yishuang, Sheng He, Zhiyong Wu, Chunxiao Xing and Liang-Jie Zhang. ‘A review of deep learning based speech synthesis’. *Applied Sciences* 9, no. 19 (2019): 4050. <https://doi.org/10.3390/app9194050>.
- Nouno, Gilbert, Arshia Cont, Grégoire Carpentier and Jonathan Harvey. ‘Making an orchestra speak’. In *Sound and music computing*. 2009. Accessed 25 February 2026. <https://inria.hal.science/hal-00839067>.
- O’Shaughnessy, Douglas. ‘Review of analysis methods for speech applications’. *Speech Communication* 151 (2023): 64–75. <https://doi.org/10.1016/j.specom.2023.05.008>.

- Oehring, Helmut. *[im'ferno] (from: MAPPA) Contrapasso I – V (to: Wladimir Putin / Sergej Lawrow)*, 2022. Accessed 25 February 2026. <https://helmutoehring.de/kopie-von-2021-22>.
- Ondrejovic, Slavo. 'Wolfgang von Kempelen and His Speaking Machine'. *Human Affairs* 2, no. 2 (1992): 161–172. <https://doi.org/10.1515/humaff-1992-020208>.
- Ore, Cecilie. *Etapper*, 1988. Accessed 25 February 2026. <https://cecilieore.no/node/92>.
- Panayotov, Vassil, Guoguo Chen, Daniel Povey and Sanjeev Khudanpur. 'Librispeech: an asr corpus based on public domain audio books'. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 5206–5210. IEEE, 2015.
- Papakyriakopoulos, Orestis, Simon Hegelich, Juan Carlos Medina Serrano and Fabienne Marco. 'Bias in word embeddings'. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 446–457. 2020. <https://doi.org/10.1145/3351095.3372843>.
- Parkinson, Adam and Graham Dunning. 'Improvising with Machines: A taxonomy of musical interactions'. *Organised Sound* 29, no. 3 (2024): 340–351. <https://doi.org/10.1017/s1355771824000268>.
- Parra, Hector. *Fragments on Fragility*, 2009. Accessed 25 February 2026. <https://hectorparra.net/docs/program-notes/>.
- Pennington, Jeffrey, Richard Socher and Christopher D. Manning. 'GloVe: Global Vectors for Word Representation'. In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. 2014. Accessed 25 February 2026. <http://aclweb.org/anthology/D14-1162>.
- Pigott, Jon. 'Across Fields: Sound, art and technology from an electromechanical perspective'. *Organised Sound* 22, no. 2 (2017): 276–285. <https://doi.org/10.1017/s1355771817000188>.
- Polashek, Timothy D. 'Beyond Babble: A Text-Generation Method and Computer Program for Composing Text, Music and Poetry'. *Leonardo Music Journal* 15 (2005): 17–22. <https://doi.org/10.1162/lmj.2005.15.1.17>.
- Prieto, Eric. 'Speech Melody and the Evolution of the Minimalist Aesthetic in Steve Reich's The Cave'. *Circuit* 12, no. 2 (2002): 21–44. <https://doi.org/10.7202/902250ar>.
- Privato, Nicola and Thor Magnusson. 'Querying the Ghost: AI Hauntography in NIME'. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, 432–438. 2024. <https://doi.org/10.5281/zenodo.13904901>.
- Puckette, Miller et al. 'Pure data: another integrated computer music environment'. In *Proceedings of the second intercollege computer music concerts*, 37–41. 1996. Accessed 25 February 2026. <https://msp.ucsd.edu/Publications/icmc97.pdf>.
- Quadrature. *Fantasia 4*, 2021. Accessed 25 February 2026. <https://quadrature.co/work/fantasi e4/>.
- Quadrature. *Noise Signal Silence*, 2019. Accessed 25 February 2026. <https://quadrature.co/work/noise-signal-silence/>.
- Ramsay, Gordon J. 'Mechanical speech synthesis in early talking automata'. *Acoustics Today* 15, no. 2 (2019): 11–19. <https://doi.org/10.1121/at.2019.15.2.11>.
- Rao, K Sreenivasa, Shashidhar G Koolagudi and Ramu Reddy Vempada. 'Emotion recognition from speech using global and local prosodic features'. *International journal of speech technology* 16 (2013): 143–160. <https://doi.org/10.1007/s10772-012-9172-2>.

- Rebguns, Maxim. *Wonderwords*, 2020. Accessed 25 February 2026. <https://github.com/mrmaxguns/wonderwordsmodule>.
- Redhead, Lauren. 'Language as a Material Process'. *Contemporary Music Review* 39, no. 3 (2020): 359–372. <https://doi.org/10.1080/07494467.2020.1821527>.
- Redhead, Lauren. *Séo níedhæmestre; se tidfara*, 2018. Accessed 25 February 2026. <https://panyrosasdiscos.bandcamp.com/track/s-o-n-edh-mestre-se-tidfara>.
- Richards, John. 'Getting the hands dirty'. *Leonardo Music Journal* 18 (2008): 25–31. <https://doi.org/10.1162/lmj.2008.18.25>.
- Richards, John and Tim Shaw. 'Improvisation through Performance-installation'. *Organised Sound* 27, no. 2 (2022): 144–155. <https://doi.org/10.1017/S1355771821000583>.
- Riches, Martin. *Talking Machine*, 1989. Accessed 25 February 2026. <https://martinriches.de/>.
- Russell, James A and Albert Mehrabian. 'Evidence for a three-factor theory of emotions'. *Journal of research in Personality* 11, no. 3 (1977): 273–294. [https://doi.org/10.1016/0092-6566\(77\)90037-x](https://doi.org/10.1016/0092-6566(77)90037-x).
- Ruviaro, Bruno. *The Spell of Speech*, 2004. Accessed 25 February 2026. <https://scholarcommons.scu.edu/music/20>.
- Sadd, Lottie. *To Copy the Speech without Saying It Directly*, 2021. Accessed 25 February 2026. <https://thebookroom.net/here-here-on-voice-and-electronics-territorialise-deterritorialise-reterritorialise/>.
- Sarhan, François. *L'Nfer*, 2012. Accessed 25 February 2026. <https://francoissarhan.blogspot.com/search?q=L%27Nfer>.
- Savery, Anna and Richard Savery. 'Collaboration Between Robots, Interfaces and Humans: Practice-Based and Audience Perspectives'. *arXiv preprint arXiv:2407.16966*, 2024. <https://doi.org/10.48550/arXiv.2407.16966>.
- Savery, Richard, Lisa Zahray and Gil Weinberg. 'Shimon sings-robotic musicianship finds its voice'. In *Handbook of Artificial Intelligence for Music: Foundations, Advanced Approaches, and Developments for Creativity*, 823–847. Springer, 2021.
- Schmucki, Annette and Alistair Zaldua. 'Music and Language Tilt into Each Other': Annette Schmucki Interviewed by Alistair Zaldua'. *Contemporary Music Review* 39, no. 3 (2020): 373–385. <https://doi.org/10.1080/07494467.2020.1821528>.
- Schroeder, Franziska and Federico Reuben. 'Breathing Life into AI: Querying AI through Tactile Imprints Using Breath and Sound'. In *Artificial Intelligence in the Cultural and Creative Sectors*, 37–54. Routledge, 2025. <https://doi.org/10.4324/9781003509738-5>.
- Schwenger, Peter. *Asemic: The art of writing*. University of Minnesota Press, 2019.
- Segnini, Rodrigo and Bruno Ruviaro. 'Analysis of Electroacoustic Works with Music and Language Intersections'. In *ICMC 2005*, 5v. 2005. Accessed 25 February 2026. <http://hdl.handle.net/2027/spo.bbp2372.2005.156>.
- Shannon, Claude Elwood. 'A mathematical theory of communication'. *The Bell system technical journal* 27, no. 3 (1948): 379–423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>.

- Sharma, Garima, Kartikeyan Umapathy and Sridhar Krishnan. ‘Trends in audio signal feature extraction methods’. *Applied Acoustics* 158 (2020): 107020. <https://doi.org/10.1016/j.apacoust.2019.107020>.
- Shepardson, Victor, Jack Armitage and Thor Magnusson. ‘Notochord: a flexible probabilistic model for real-time MIDI performance’. *arXiv preprint arXiv:2403.12000*, 2024. <https://doi.org/10.48550/arXiv.2403.12000>.
- Shriberg, Elizabeth, Raj Dhillon, Sonali Bhagat, Jeremy Ang and Hannah Carvey. ‘The ICSI meeting recorder dialog act (MRDA) corpus’. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, 97–100. 2004. <https://doi.org/10.21236/ada460980>.
- Sivertsen, Christian, Maarten Smith and Sander Van Der Zwan. ‘Art Critique by Other Means’. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, 343–352. 2023. <https://doi.org/10.1145/3563657.3596069>.
- Smalley, Denis. ‘Defining Transformations’. *Interface* 22, no. 4 (1993): 279–300. <https://doi.org/10.1080/09298219308570638>.
- Smith, Benjamin D and W Scott Deal. ‘ML.* Machine learning library as a musical partner in the computer-acoustic composition flight’. In *Proceedings of the 2014 International Computer Music Conference*, 1285–1289. Michigan Publishing, 2014. Accessed 25 February 2026. <https://hdl.handle.net/1805/15340>.
- Spahn, Wolfgang. *Echo State Network*, 2019. Accessed 25 February 2026. http://wolfgang-spahn.com/doku.php/installation:echo_state_networks.
- Stapleton, Paul and Tom Davis. ‘Ambiguous Devices: Improvisation, agency, touch and feed-through in distributed music performance’. *Organised Sound* 26, no. 1 (2021): 52–64. <https://doi.org/10.1017/S1355771821000054>.
- Stolcke, Andreas, Klaus Ries, Noah Cocco, Elizabeth Shriberg, Rebecca Bates, Dan Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema and Marie Meteer. ‘Dialogue act modeling for automatic tagging and recognition of conversational speech’. *Computational linguistics* 26, no. 3 (2000): 339–374.
- Swain, Monorama, Aurobinda Routray and Prithviraj Kabisatpathy. ‘Databases, features and classifiers for speech emotion recognition: a review’. *International Journal of Speech Technology* 21 (2018): 93–120. <https://doi.org/10.1007/s10772-018-9491-z>.
- Tan, Xu, Tao Qin, Frank Soong and Tie-Yan Liu. ‘A Survey on Neural Speech Synthesis’. *arXiv*, 2011, 2106.15561. <https://doi.org/10.48550/arXiv.2106.15561>. eprint: 2106.15561.
- Team, Magenta. *Magenta: Music and Art Generation with Machine Intelligence*, 2016. Accessed 25 February 2026. https://github.com/magenta/magenta/tree/main/magenta/models/polyphony_rnn.
- Tenney, James. *Ain’t I a Woman?*, 1992. Accessed 25 February 2026. <https://brahms.ircam.fr/en/works/work/46265/>.
- Thompson IV, W. A. ‘Creating Musical Scores Inspired by the Intersection of Human Speech and Music Through Model-Based Cross Synthesis’. PhD diss., Louisiana State University, 2022.

- Tremblay, Pierre Alexandre, Owen Green, Gerard Roma, James Bradbury, Ted Moore, Jacob Hart and Alex Harker. *The Fluid Corpus Manipulation Toolbox (v.1)*. V. v1.0.0, 2022. <https://doi.org/10.5281/zenodo.6834643>.
- Tremblay, Pierre Alexandre, Gerard Roma and Owen Green. ‘Enabling programmatic data mining as musicking: The fluid corpus manipulation toolkit’. *Computer Music Journal* 45, no. 2 (2021): 9–23. https://doi.org/10.1162/comj_a_00600.
- Truax, Barry. *Pendlerdrøm*, 1997. Accessed 25 February 2026. <https://electrocd.com/en/oeuvre/13157/barry-truax/pendlerdrm>.
- Truax, Barry. ‘Speech, music, soundscape and listening: interdisciplinary explorations’. *Interdisciplinary Science Reviews* 47, no. 2 (2022): 279–293. <https://doi.org/10.1080/03080188.2022.2035103>.
- Valle, Andrea. ‘Making acoustic computer music: The Rumentarium project’. *Organised Sound* 18, no. 3 (2013): 242–254. <https://doi.org/10.1017/s1355771813000216>.
- Van Den Oord, Aaron, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu et al. ‘Wavenet: A generative model for raw audio’. *arXiv preprint arXiv:1609.03499* 12, no. 1 (2016). <https://doi.org/10.48550/arXiv.1609.03499>.
- Vassallo, J. *Versificator*, 2021. Accessed 25 February 2026. <https://nivel.teak.fi/carpa7/versificator-algorithmic-poetry-and-music-composition/>.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser and Illia Polosukhin. ‘Attention is all you need’. *Advances in neural information processing systems* 30 (2017).
- Vear, Craig, Steve Benford, Juan Martinez Avila and Solomiya Moroz. ‘Human-AI Musicking: A Framework for Designing AI for Music Co-creativity’. *AIMC 2023*, 2023. Accessed 25 February 2026. <https://aimc2023.pubpub.org/pub/zd46ltn3>.
- Venables, Phillip. *Answer Machine Tape, 1987*, 2022. Accessed 25 February 2026. <https://philipvenables.com/2022/09/25/answer-machine-tape-1987>.
- Ververidis, Dimitrios and Constantine Kotropoulos. ‘Emotional speech recognition: Resources, features, and methods’. *Speech communication* 48, no. 9 (2006): 1162–1181. <https://doi.org/10.1016/j.specom.2006.04.003>.
- Villicaña-Shaw, Nathan, Spencer Salazar and Ajay Kapur. ‘Mechatronic Performance in Computer Music Compositions.’ In *Proceedings of the International Conference on New Interfaces for Musical Expression*, 413–418. 2018. <https://doi.org/10.5281/zenodo.1302647>.
- Walshe, Jennifer. *IS IT COOL TO TRY HARD NOW?*, 2017. Accessed 25 February 2026. <https://milker.org/jenniferwalshe-worklist>.
- Walshe, Jennifer. *The Text Score Dataset 1.0*, 2021. Accessed 25 February 2026. <http://milker.org/s/The-Text-Score-Dataset-10-Jennifer-Walshe.pdf>.
- Wang, Fali, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, Tzuhao Mo, Qiuha Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang et al. ‘A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness’. *ACM Transactions on Intelligent Systems and Technology* 16, no. 6 (2025): 1–87. <https://doi.org/10.1145/3768165>.

- Wang, Lei, Ziyi Zhao, Hanwei Liu, Junwei Pang, Yi Qin and Qidi Wu. 'A review of intelligent music generation systems'. *Neural Computing and Applications* 36, no. 12 (2024): 6381–6401. <https://doi.org/10.1007/s00521-024-09418-2>.
- Wingfield, Paul. 'Janáček's speech-melody theory in concept and practice'. *Cambridge Opera Journal* 4, no. 3 (1992): 281–301. <https://doi.org/10.1017/s0954586700003803>.
- Wishart, Trevor. *Globalalia*, 2004. Accessed 25 February 2026. <https://www.soundohm.com/product/globalalia-imago>.
- Wishart, Trevor. 'The Composition of "Vox-5"'. *Computer Music Journal* 12, no. 4 (1988): 21–27. <https://doi.org/10.2307/3680150>.
- Wishart, Trevor. *Vox 5*, 1986. Accessed 25 February 2026. <https://britishmusiccollection.org.uk/score/vox-5>.
- Yoshimura, Fou and Kazuhiro Jo. 'A "voice" instrument based on vocal tract models by using soft material for a 3D printer and an electrolarynx.' In *Proceedings of the International Conference on New Interfaces for Musical Expression*, 411–412. 2019. <https://doi.org/10.5281/zenodo.3673015>.
- Young, John. *An Angel at Mons*, 2014. Accessed 25 February 2026. <https://empreintesdigitales.bandcamp.com/track/an-angel-at-mons>.
- Young, John. 'Figures of Speech: Oral History as an Agent of Form in Electroacoustic Music'. *Leonardo Music Journal* 28 (2018): 88–94. https://doi.org/10.1162/lmj_a_01047.
- Z, Pamela. *Geekspeak*, 1995. Accessed 25 February 2026. <https://pamelaz.com/work/geekspeak/>.
- Zareei, Mo H. 'Audiovisual materialism'. *Organised Sound* 25, no. 3 (2020): 362–371. <https://doi.org/10.1017/s1355771820000321>.
- Zareei, Mo H, Dale A Carnegie and Ajay Kapur. 'Noise Square: Physical Sonification of Cellular Automata through Mech-atronic Sound-sculpture'. In *Proceedings of the 21st International Symposium on Electronic Art*. Vancouver, Canada, 2015. Accessed 25 February 2026. https://www.isea-archives.org/docs/2015/proceedings/ISEA2015_proceedings.pdf.
- Zemanek, Heinz. 'Geschichte des Automaten: Das Componium von Winkel Ein Gerät mit einem Zufallsgenerator aus dem Jahre 1821'. *it-Information Technology* 8, nos. 1-6 (1966): 61–62. <https://doi.org/10.1524/itit.1966.8.16.61>.
- Zuiderveld, J. *Conversations beyond the ordinary*, 2024. Accessed 25 February 2026. <https://ars.electronica.art/hope/en/conversations-beyond-the-ordinary/>.