

**Place, authenticity, and time:
a framework for
liveness in synthetic speech**

Christopher H. Newell

Submitted for the degree of Doctor of Philosophy

The University of York

The Human-Computer Interaction Group

The Department of Computer Science

September 2009

Abstract

In his 'Uncanny Valley', Mori posits that, when a non-human entity appears too 'realistic' or human-like, users respond with unease. To date, most synthetic speech research has ignored the phenomenon of the 'Uncanny Valley', striving for greater 'realism' and verisimilitude to human speech. Distinct from 'realism', however, is the concept of 'liveness', which is more akin to 'spontaneity' or 'truthfulness'. This thesis documents the development of the Place, Authenticity, Time (PAT) framework, a framework that seeks to support the design of synthetic voices by making them appear more 'live', rather than more 'realistic'. The framework is derived both from the performing arts and from computer science, and casts synthetic voices as 'synthetic voice actors'. Over the course of this thesis, a software system is built and a number of studies are undertaken to reveal whether simple, performing-arts-derived, paralinguistic/prosodic modifiers can improve the user perception of a synthetic voice's 'acting'. These studies show a preference for a modified synthetic voice over a non-modified synthetic voice in straightforward head-to-head comparisons, but no clear preference in more complex interactive implementations. The PAT framework emerges from these studies as a model of the 'problem space' confronting the designer of synthetic voice actors. It suggests ways in which acting, scripting and setting can be used to evoke 'liveness', and to sidestep issues of 'realism' and the accompanying user unease.

List of contents

1	INTRODUCTION.....	1
1.1.1	Discussion.....	1
1.1.2	Evolution of the research.....	3
1.1.3	What is the PAT framework?	4
1.1.4	Text-to-speech synthesis and the framework.....	7
1.1.5	The current state of speech synthesis	8
1.1.6	Realism v the willing suspension of disbelief	12
1.1.7	Liveness	13
1.1.8	Place, Authenticity and Time	15
1.1.9	Testing the PAT framework.....	16
1.1.10	The propositions examined in this thesis	18
1.1.11	Defining sources, perspectives and features	18
1.1.12	Methodologies applied in this research	19
1.1.13	Constraints applied to this research	22
1.2	Chapter Summary.....	24
1.2.1	Chapter 1: Introduction.....	24
1.2.2	Chapter 2: Sources, perspectives and features for the PAT framework	24
1.2.3	Chapter 3: The language of speech related phenomena	24
1.2.4	Chapter 4: Detecting and describing liveliness and liveness.....	25
1.2.5	Chapter 5: Automated paralinguistic/prosodic modification	25
1.2.6	Chapter 6: Evaluating the values for the paralinguistic/prosodic modifiers.....	25
1.2.7	Chapter 7: Evaluating the PAT framework.....	26
1.2.8	Chapter 8: Potential applications.....	26
1.2.9	Chapter 9: Conclusion and future research	26
1.3	Conclusion.....	27
2	SOURCES, PERSPECTIVES AND FEATURES FOR THE PAT FRAMEWORK ..	28
2.1	Breadth and depth	28
2.2	Additive, distractive or subtractive processes	29
2.3	Artificial intelligence and robotics.....	33
2.4	Linguistics - prosody	35
2.4.1	Paralanguage.....	37
2.4.2	Spontaneous speech	38
2.5	Music, the performing arts and cinema.....	39
2.5.1	What is acting?.....	39
2.5.2	Finding rules for synthetic acting	41
2.5.3	Renaissance verse speaking	45

2.5.4	Ventriloquism.....	48
2.5.5	Singing and opera.....	50
2.5.6	Extended vocal technique.....	52
2.5.7	Musical expression.....	54
2.5.8	Humanisation algorithms and randomness.....	56
2.5.9	Liveness.....	61
2.5.10	Cinematic sounds.....	64
2.6	Metaphysical reflections on silence.....	69
2.7	A theoretical basis for the PAT tool.....	72
2.8	A theoretical basis for the PAT framework.....	75
2.9	Conclusions.....	79
3	THE LANGUAGE OF SPEECH RELATED PHENOMENA.....	80
3.1	Illustrating the problem of ambiguous terminology.....	82
3.2	Addressing the problem of ambiguous terminology.....	83
3.2.1	The structured vocabulary.....	84
3.3	The key speech terms for this thesis.....	86
3.3.1	Neutral Speech.....	86
3.3.2	Speech.....	86
3.3.3	Standard speech.....	87
3.3.4	Voice.....	87
3.3.5	A character voice.....	88
3.3.6	An affective voice.....	89
3.3.7	Speaking and reading: a discussion.....	90
3.4	Conclusions.....	91
4	DETECTING AND DESCRIBING LIVELINESS AND LIVENESS.....	93
4.1	Liveliness terminology test.....	94
4.2	Hearing liveliness test.....	96
4.3	Conclusions.....	102
5	AUTOMATED PARALINGUISTIC/PROSODIC MODIFICATION.....	104
5.1	Paralinguistic/prosodic modifiers: perspectives from linguistics.....	106
5.1.1	Pauses.....	107

5.1.2	Breaths and filled pauses in spontaneous speech	109
5.1.3	Periodic tempo (speech rate) variations	110
5.1.4	A summary of pauses from the linguistic literature.....	112
5.1.5	Prosody in synthetic speech is a complexity problem: a discussion	113
5.2	Paralinguistic/prosodic modifiers	115
5.3	Prosodic representation and mark-up.....	116
5.3.1	Automated mark-up in the PAT software tool.....	120
5.3.2	Randomness in automated mark-up.....	121
5.3.3	Filtered randomness in automated mark-up	121
5.4	Conclusions	124
6	EVALUATING THE VALUES FOR THE PARALINGUISTIC/PROSODIC MODIFIERS.....	126
6.1.1	Application environment	126
6.1.2	Application functions in the PAT software tool	127
6.1.3	The Microsoft default settings	135
6.1.4	PAT Software tool functions subject to user value assignment.....	138
6.1.5	Methodology for assigning values to the paralinguistic/ prosodic modifiers.....	139
6.1.6	Evaluating values for the prosodic/paralinguistic modifiers - focus group.....	139
6.1.7	Evaluating values for the paralinguistic/prosodic modifiers - Web based pilot study (1)	142
6.1.8	Evaluating values for the paralinguistic/prosodic modifiers - Web based pilot study (2)	149
6.1.9	Evaluating values for the paralinguistic/prosodic modifiers - User survey.....	157
6.1.10	Evaluating the perception of 'liveness' - Web-based pilot study (3)	162
6.2	Setting the values for the prosodic modifiers	169
6.2.1	Empty pauses at natural punctuation points.....	169
6.2.2	Empty non-grammatical pauses.....	171
6.2.3	Breath filled pauses.....	171
6.2.4	Periodic tempo (speech rate) variations	172
6.2.5	Background sounds.....	172
6.2.6	Choice of texts.....	173
6.3	Conclusions	173
7	EVALUATING THE PAT FRAMEWORK	175
7.1	Overview.....	175
7.1.1	The design of the PAT framework tests	177
7.1.2	'Tide'.....	179
7.1.3	'Call Centre'.....	183
7.1.4	'Please wait with me	184
7.1.5	'PAT Testing:' testing the Framework	192

7.2	Can liveness be quantified?	205
7.3	Conclusions	208
8	POTENTIAL APPLICATIONS	210
8.1	Telephone banking	211
8.2	The technology museum.....	215
8.3	A toy polar bear	218
8.4	Conclusions	220
9	CONCLUSION AND FUTURE RESEARCH.....	221
9.1	Conclusions to the three propositions.....	221
9.2	Conclusions to Proposition 1.....	222
9.3	Conclusions to Proposition 2.....	223
9.4	Conclusions to Proposition 3.....	224
9.5	General conclusions.....	224
9.5.1	Addressing the title of the thesis	226
9.5.2	The contribution of this research.....	227
9.6	Further research.....	231
9.6.1	The composition of prosodic modifiers	231
9.6.2	Inter-industry development.....	232
9.6.3	Deeper ontologies.....	232
9.6.4	Expressive speech without text	233
9.6.5	Breath-like pauses with sounds other than breaths	235
9.6.6	Further tests to demonstrate the predictive powers of the framework	235
9.6.7	Creative Speech Technology Network – CreST Network.	235
9.6.8	Formalising texts for comparative testing of synthesis systems.....	236
9.6.9	Methodologies derived from elocution	236
9.7	Finale	236
	APPENDIX.....	237
	BIBLIOGRAPHY	271

List of Tables

Table 1: Table showing the Place, Authenticity, Time (PAT) framework.....	16
Table 2: Defining sources, perspectives, features and metrics	19
Table 3: Evolution of the PAT framework and the methodologies applied.....	21
Table 4: Symbols for paralanguage.....	37
Table 5: Perspectives and features to inform the theoretical basis for the PAT software tool ..	74
Table 6: Features encoded in the PAT software tool.....	75
Table 7: Structured vocabulary for evaluating synthetic speech.....	85
Table 8: Conceptual groups defining liveliness.....	95
Table 9: Categories of properties for liveliness	96
Table 10: Frequency for identification of recordings 1 to 3 and tone phrases 1 to 3	100
Table 11: Results from test 1: hearing liveliness test – speech. The P column shows the p-value for a chi-square test of this data as reported by Excel software.	101
Table 12: Results from test 2: hearing liveliness test – tones. The P column shows the p-value for a chi-square test of this data as reported by Excel software.	101
Table 13: Breath pauses and non-breath pauses in spontaneous speech and readings. From a study by Goldman Eisler.	109
Table 14: Breath pauses at grammatical and non-grammatical junctures. From a study by Goldman Eisler.	110
Table 15: Linguistic components that can be represented as plain text	118
Table 16: The list of prosodic and paralinguistic variables considered in this research.....	119
Table 17: The specification for a filed pause with breath sound in the PAT software tool.....	135
Table 18: The results of tests specifying the reliability of the PAT software tool.....	136
Table 19: Microsoft Mary compared to a human speaker	137
Table 20: Settings available for user modification in the PAT tool.....	138
Table 21: Results from the focus group to set values for prosodic modifiers	141
Table 22: Setting ranges applied to the prosodic and paralinguistic modifiers	144
Table 23: Results of web-based pilot study (1). The P column shows the p-value for a chi-square test of this data as reported by Excel software.	146
Table 24: Results of any pause edits compared to no pause edits. The P column shows the p-value for a chi-square test of this data as reported by Excel software.	148
Table 25: Description of audio recordings for web based pilot study (2).....	155
Table 26: Results of tests 1, 2 and 4 for web based pilot study (2). The P column shows the p-value for a chi-square test of this data as reported by Excel software.	156
Table 27: Results of test 3 for web based pilot study (2). P shows the p-value for a chi-square test of this data as reported by Excel software.	156
Table 28: Results of the user survey. The P column shows the p-value for a chi-square test of this data as reported by Excel software.....	161
Table 29: Results of web-based pilot study (3). The P row shows the p-value for a chi-square test of this data as reported by Excel software. Incorrect answers are not included in the calculation of the chi-square test.	168

Table 30: The recordings used in ‘Tide’	180
Table 31: Results from the ‘please wait with me’ installation at the university campus	191
Table 32: Results from the ‘please wait with me’ installation at a digital arts festival	191
Table 33: Results of the ‘PAT Testing’ performances. The result of the Friedman Test as reported by the XLSTAT software is shown in the bottom row.....	201
Table 34: Histograms of the results of the ‘PAT Testing’ performances. The x axis shows the score measured in mm: the y axis shows the number of participants.....	202
Table 35: Performance 3/5 scored for liveness	207
Table 36: Performance 6/9 scored for liveness	208
Table 37: The PAT framework applied to a telephone banking voice	215
Table 38: The PAT framework applied to a museum guide voce	218
Table 39: The PAT framework applied to the voice of a toy polar bear	220
Table 40: The screens providing information for the audience in ‘PAT Testing’	270

List of Figures

Figure 1: Von Kempelen’s “synthesizer” (image from (Black 2009)), Sparky’s Magic Piano (image from (Flickr 2009)) and HAL (Kubrick & Clarke 1968)	2
Figure 2: The context for the PAT framework	6
Figure 3: Professor Hiroshi Ishiguro (standing) and Repliee Q1Expo (sitting)	11
Figure 4: ‘The Uncanny Valley.’ (Mori 1970)	35
Figure 5: Samuel Beckett’s notes for pauses in ‘Not I’ (Gontarski 1992).....	43
Figure 6: A line of iambic pentameter. Weak beats are indicated by ‘.’ and strong beats by ‘_’	46
Figure 7: Plotting emotions to musical mechanisms from (Juslin & Sloboda p.315 ibid.).....	55
Figure 8: Waveform for the human performance	58
Figure 9: Waveform and notation for the machine performance	58
Figure 10: Waveform and notation for the “humanized” machine performance.....	59
Figure 11: An aspirational synthetic speech system.....	76
Figure 12: Current speech system outcomes	77
Figure 13: An alternative speech system	77
Figure 14: A “visualisation” of the PAT framework	78
Figure 15: The game ‘Moods’ (©Hasbro Inc).....	90
Figure 16: Sample answer sheet for the hearing liveliness test	99
Figure 17: Temporal patterns in speech. A: reading; B: spontaneous speech.	111
Figure 18: 300 random bits generated by the C++ function <code>rand() % 2</code>	122
Figure 19: 300 filtered random bits	122
Figure 20: Subjective randomness.....	123
Figure 21: Filtered random integers according to Rabin’s rules.....	124
Figure 22: Schematic of the PAT software tool	129
Figure 23: The interface for the PAT software tool	134
Figure 24: Interface for the web-based pilot study (1).....	146
Figure 25: The frequency of edits for each of the three categories of pauses.....	147
Figure 26: Instruction screen for web based pilot study (2).....	151
Figure 27: Introductory screen for web based pilot study (2)	152
Figure 28: Test 1 for web based pilot study (2)	152
Figure 29: Test 2 for web based pilot study (2)	153
Figure 30: Test 3 for web based pilot study (2)	153
Figure 31: Test 4 and summary dialogue box for web based pilot study (2).....	154
Figure 32: Instruction page for user survey	159
Figure 33: Sample answer page for user survey	160
Figure 34: Introduction screen for web-based pilot study (3).....	164
Figure 35: ‘Play’ screen for web-based pilot study (3)	165
Figure 36: Participant details screen for web based pilot study (3)	165
Figure 37: Questions screen for web based pilot study (3)	166
Figure 38: Data collected from web-based pilot study (3)	167
Figure 39: A composite image of the ‘Tide’ installation	181

Figure 40: The 'Tide' artwork on display.....	182
Figure 41: The adapted old-fashioned telephone used in 'Call Center.'	183
Figure 42: Post-card advertising the 'please wait with me' installation.....	185
Figure 43: The 'please wait with me' installation in use in a university corridor.	189
Figure 44: The 'please wait with me' installation at a digital arts festival.....	190
Figure 45: A postcard advertising the 'PAT Testing' performance.	194
Figure 46: A still frame of the mouth used in 'Not I' reprise	199
Figure 47: Chart showing the comparative truth rating of the 'PAT Testing' performances. The x axis shows the performance reference. The y axis shows the total score in mm.	203
Figure 48: A proposed direction for alternative speech research	234
Figure 49: 'PAT Testing' audience survey report card. Page 1.....	254
Figure 50: 'PAT Testing' audience report card. Page 2	255
Figure 51: Page 1 of the programme notes for 'PAT Testing'	256
Figure 52: Page 2 of the programme notes for 'PAT Testing'	257
Figure 53: Page 3 of the programme notes for 'PAT Testing'	258
Figure 54: Page 4 of the programme notes for 'PAT Testing'	259
Figure 55: Page 5 of the programme notes for 'PAT Testing'	260
Figure 56: Stage plan and equipment set up for 'PAT Testing'	261

Acronyms

AI	Artificial Intelligence
CreST	Creative Speech Technology
EVT	Extended Vocal Technique
HCI	Human-Computer Interaction
MIDI	Musical Instrument Digital Interface
PAT	Place Authenticity Time
SALT	Speech Application Language Tags
SAPI	Speech Application Programming Interface
SDK	Software Development Kit
SSML	Synthetic Speech Mark-up Language
TTS	Text-to-Speech Synthesis
WSOD	Willing Suspension of Disbelief

DVD Tracks

DVD 1 contains an audiovisual summary of this research. It may provide a useful introduction to the thesis. The running time is 14 minutes.

DVD 2 contains the complete 'PAT Testing' performance. It is included for completeness' sake and does not provide a pleasing viewing experience. It is shot with a single static camera under very low light levels with negligible visual material of interest. Viewers are advised either not to watch it at all or to listen to it with the video image switched off. The running time is 72 minutes 39 seconds.

Audio CD Tracks

The **AUDIO CD** contains tracks referenced in the body of the thesis. The file names are the titles as may be displayed on some types of CD player.

Track 1 01_sentence.aiff

York Talk: University of York

<http://www.ims.uni-stuttgart.de/~moehler/synthspeech/#english>

Track 2 02_diphone.wav

Festival Speech System: CSTR Edinburgh

<http://www.ims.uni-stuttgart.de/~moehler/synthspeech/#english>

Track 3 03_unitsel.wav

Festival Speech System: CSTR Edinburgh

<http://www.ims.uni-stuttgart.de/~moehler/synthspeech/#english>

Track 4 04_wduis6.aiff

Haskins Laboratories: Yale University

<http://www.haskins.yale.edu/facilities/DYNAMIC/sentence.html>

Track 5 05_Simon.wav

Loquendo Vocal Technology and Services

<http://tts.loquendo.com/ttsdemo/default.asp?page=id&language=en>

Track 6 6_Dodge_1.wav

Dodge, C. (2004) 'Speech Songs (1973 – Excerpt: No. 1. When I Am With You)' in Amirkhanian, C., Anderson, B., Ashley, R., et al. *10 + 2: 12 American Text Sound Pieces* (audio recording on compact disc). San Francisco: Other Minds.
(Extract)

Track 7 7_fitter_happier.wav``

(Extract)

Track 8 8_h_harry_ich_bin_nich_glueklieh.wav

Huge Harry, (MIT, various authors, 1992 - 2000)

Excerpt from Luuk Bouwman's film 'Huge Harry and the Institute of Artificial Art.'

<http://www.iaaa.nl/hh/LBclips/sound-win.html>

- Track 9** 9_benassi_selection.wav
Benassi, B. (2003) 'Satisfaction' in Benassi, B. *Hypnotica* (audio recording on compact disc). New York City: Ultra Records.
(Extract)
- Track 10** 10_wright_linney_sample1_190607.wav
Wright & Linney, Conversation piece, 2007
<http://www.ucl.ac.uk/conversation-piece>
(Extract)
- Track 11** 11_MADAMEi.wav
Hood, Madame I, 2007
<http://www.madamei.com/home.htm>
(Extract)
- Track 12** 12_tosca_live.wav
Puccini, G. (2000) 'Vittoria! Vittoria!' in Puccini, G. *Tosca*. Performed by Corelli, F., tenor (audio recording on compact disc). Milford, Connecticut: Bel Canto Society.
(Extract)
- Track 13** 13_tosca_studio.wav
Puccini, G. (2007) 'Nel pozzo, nel giardino' in Puccini, G. *Tosca*. Performed by Corelli, F., tenor (audio recording on compact disc). Milan: Urania.
(Extract)
- Track 14** 14_r2d2wst3.wav
Kurtz, G., Lucas, G., McCallum, R. (1977) *Star Wars Episode IV: A New Hope* (DVD). 20th Century Fox.
- Track 15** 15_the_sims2_111.mp3
Simlish - The Sims – Maxis/Electronic Arts 2000,
Voice actors Stephen Kearin and Gerri Lawlor.
- Track 16** 16_msnd_hall.wav
A Midsummer Night's Dream. Shakespeare, W., Directed by Peter Hall, Broadcast for Television 1966
(Extract) 2.2.151
- Track 17** 17_msnd_bbc.wav
A Midsummer Night's Dream. Shakespeare, W., BBC Shakespeare. London, BBC Worldwide, DVD, 2005
(Extract) 2.2.151
- Track 18** 18_msnd_RSC.wav
A Midsummer Night's Dream. Shakespeare, W., RSC., Directed by Adrian Nobel, Film Four, DVD, 1996
(Extract) 2.2.151
- Track 19** 19_msnd_HOLLYWOOD.wav
A Midsummer Night's Dream. Shakespeare, W., Directed by Hoffman, M., Fox Searchlight, DVD, 1999.
(Extract) 2.2.151

Track 20 20_msnd_REINHARDT.wav

A Midsummer Night's Dream. Shakespeare, W., Warner Brothers, DVD, 1935.
(Extract) 2.2.151

Track 21 21_accompagnato.wav

Telemann, G. (1999) 'Recitativo accompagnato: Denn der Herr horet mein Weinen' from 'Psalm 6: Ach Herr, strafe mich nicht, (TWV 7:1)' in Buxtehude, D., Bach, J., Telemann, G. *German Church Cantatas and Arias*. Performed by Jacobs, R., counter-tenor, and the Kuijken Consort (audio recording on compact disc). Heidelberg: Accent Records.
(Extract)

Track 22 22_misurato.wav

Telemann, G. (1999) 'Recitativo misurato' from 'Ihr Volker, hort, (TWV 1:921)' in Buxtehude, D., Bach, J., Telemann, G. *German Church Cantatas and Arias*. Performed by Jacobs, R., counter-tenor, and the Kuijken Consort (audio recording on compact disc). Heidelberg: Accent Records.
(Extract)

Track 23 23_cosi.wav

Mozart, W. (2008) 'Act I: Recitative: Swords or pistols?' in Mozart, W. *Così fan tutte* (K. 588). Conducted by Sir Charles Mackerras, performed by the Orchestra of the Age of Enlightenment (audio recording on compact disc). Colchester: Chandos.
(Extract)

Track 24 24_handel_recits.wav

Handel, F. (2007) 'Part I, Nos 14a - 16: There were shepherds abiding in the field' in Handel, F. *Messiah*. Conducted by Sir Colin Davis, performed by Susan Gritton, soprano, and the London Symphony Orchestra (audio recording on compact disc). London: LSO Live.
(Extract)

Track 25 25_fledermaus_melodrama.wav

Strauss, J. (1986) 'No. 5. Finale' in Strauss, J. *Die Fledermaus*. Conducted by Plácido Domingo (audio recording on vinyl). London: EMI.
(Extract)

Track 26 26_sstimme.wav

Schoenberg, A. (1995) 'A Survivor from Warsaw, Op. 46' in Schoenberg, A. *Das Chorwerk*. Conducted by Pierre Boulez (audio recording on compact disc). New York: Sony Classical.
(Extract)

Track 27 27_bach_cello_human.wav

Bach, J. (2005) 'Suite for solo cello No. 1, in G major (BMV 1007)' in Bach, J. 'The Cello Suites'. Performed by Jian Wang, cello (audio recording on compact disc). Hamburg: Deutsche Grammophon.
(Extract)

Track 28 28_bach_unquantised.wav

- Track 29** 29_bach_time_quantised 30ms.wav
- Track 30** 30_bach_time_30ms and_velocity_50%.wav
- Track 31** 31_BERIO_edit.wav
 Berio, L., 'Sequenza III' in Berio, L. *Circles – Sequenza I – Sequenza III – Sequenza V*.
 Performed by Cathy Berberian, mezzo soprano (audio recording from compact disc).
 Mainz: Wergo.
 (Extract)
- Track 32** 32_cage.wav
 Cage, J. (1998) 'Solo for Voice 22 (from Songbooks)' in Cage, J. *Litany for the Whale*.
 Performed by Paul Hillier and Theatre of Voices (audio recording from compact disc).
 Arles: Harmonia Mundi.
 (Extract)
- Track 33** 33_pscho.wav
 Hitchcock, A., *Psycho*. London, BBC. 1960
 (Extract)
- Track 34** 34_perez.wav
 Prado, P. (2008) 'More Mambo Jambo' in Prado, P. *El Rey Del Mambo – His 27 Finest 1949 – 1956* (audio recording on compact disc). London: Retrospective.
 (Extract)
- Track 35** 35_sibelius.wav
 Sibelius, J. (1992) 'Third Movement' in Sibelius, J. *Symphony No. 5, in E flat major (Op. 82)*. Conducted by Lorin Maazel (audio recording on compact disc). London: Decca.
 (Extract)
- Track 36** 36_cockney.wav
 Steve Harley & Cockney Rebel. (1975) 'Make Me Smile (Come Up and See Me)' in Steve Harley & Cockney Rebel. *The Best Years of Our Lives* (audio recording on compact disc). London: EMI.
 (Extract)
- Track 37** 37_humphries.wav
 BBC Radio 4, Afternoon Play, 'Baring Up',
 Broadcast - Thursday 19 March, 2.15 - 3.00pm 2009
 (Extract)
- Track 38** 38_Hearing liveliness test 1
- Track 39** 39_Hearing liveliness test 2
- Track 40** 40_Hearing liveliness test 3
- Track 41** 41_Hearing liveliness test 4
- Track 42** 42_Hearing liveliness test 5
- Track 43** 43_Hearing liveliness test 6
- Track 44** 44_peston.wav
 Robert Peston, BBC Economics Correspondent, YouTube broadcast
http://news.bbc.co.uk/1/hi/programmes/politics_show/7888284.stm
 (Extract)

- Track 45** 45_Focus group original
- Track 46** 46_Focus group outcome
- Track 47** 47_Web based pilot study 1 ex 1
- Track 48** 48_Web based pilot study 1 ex 2
- Track 49** 49_Web based pilot study 1 ex 3
- Track 50** 50_Web based pilot study 1 ex 4
- Track 51** 51_Web based pilot study 2 ex 1
- Track 52** 52_Web based pilot study 2 ex 2
- Track 53** 53_Web based pilot study 2 ex 3
- Track 54** 54_Web based pilot study 2 ex 4
- Track 55** 55_Web based pilot study 2 ex 5
- Track 56** 56_Web based pilot study 2 ex 6
- Track 57** 57_Web based pilot study 2 ex 7
- Track 58** 58_Web based pilot study 2 ex 8
- Track 59** 59_Web based pilot study 2 ex 9
- Track 60** 60_Web based pilot study 2 ex 10
- Track 61** 61_Web based pilot study 2 ex 11
- Track 62** 62_Web based pilot study 2 ex 12
- Track 63** 63_Web based pilot study 2 ex 13
- Track 64** 64_Web based pilot study 2 ex 14
- Track 65** 65_Web based pilot study 3 ex 1
- Track 66** 66_Web based pilot study 3 ex 2
- Track 67** 67_User survey ex 1
- Track 68** 68_User survey ex 2
- Track 69** 69_User survey ex 3
- Track 70** 70_User survey ex 4
- Track 71** 71_User survey ex 5
- Track 72** 72_User survey ex 6
- Track 73** 73_Tide example 1
- Track 74** 74_Tide example 2
- Track 75** 75_Tide example 3
- Track 76** 76_Tide example 4
- Track 77** 77_Tide example 5
- Track 78** 78_Call center example
- Track 79** 79_Please wait with me example
- Track 80** 80_oudeyer 1
Oudeyer, P, 2004
<http://www.csl.sony.fr/~py/> Accessed: 28/10/2004
- Track 81** 81_oudeyer 2
Oudeyer, P, 2004
<http://www.csl.sony.fr/~py/> Accessed: 28/10/2004

Track 82	82_Macintalk example 1 System 10.3.9	'Bahh' - Apple Macintosh Operating
Track 83	83_Macintalk example 2 System 10.3.9	'Deranged' - Apple Macintosh Operating
Track 84	84_Macintalk example 3 System 10.3.9	'Hysterical' - Apple Macintosh Operating
Track 85	85_Macintalk example 4 System 10.3.9	'Pipe organ' - Apple Macintosh Operating
Track 86	86_Macintalk example 5 System 10.3.9	'Trinoids' - Apple Macintosh Operating
Track 87	87_Macintalk example 6 System 10.3.9	'Whisper' - Apple Macintosh Operating

Preface on interdisciplinary research

Interdisciplinarity: "... a mutual appropriation and reappropriation of tools and techniques from one field to another with no assumption of a shared understanding either at the level of problems or at the level of concerns and background." (Franchi & Güzeldere 2005, p.101)

This thesis is motivated by a curiosity to find out whether the sort of techniques employed by actors to improve their voice acting could be translated into rules for a synthesiser that would lead to better speech synthesis.

There exist no definitive, documented rules for acting. Much that has been said about acting that could potentially be turned into rules only exists verbally, as advice passed from performer to performer. Some performing arts practitioners and teachers (Stanislavskii, Brecht, Artaud and Boal come to mind) write enough about their work for them to achieve the status of theorists. Another group of practitioners don't believe in theory at all, preferring to describe the process as personal, intuitive and largely unteachable. By far the largest group are engaged in the struggle to find work and do not have time to devote to documenting the process: consequently, in the performing arts, 'theories' frequently exist for the duration of rehearsals and then disappear, resurfacing only as anecdotes. Such material cannot be included in a computer science thesis, and consequently, the culminating event of this research is a performance in which an attempt is made to capture evidence from a performance event before it is relegated to anecdote, and use it to draw conclusions relevant to computer science. To my knowledge the only other comparable interdisciplinary work is that of Alan F. Newell¹ (Newell, Carmichael, Morgan et al. 2006) and the Foxtrot Theatre Company (Foxtrot 2009) in which human-computer interaction (HCI) specialists, directors and actors work together to elicit requirements for enabling technologies for older people. This is particularly significant innovation in that it is hosted by an academic institution, with a theatre space situated in a computer science department².

¹ Alan F. Newell is not a relation of the author.

² The University of Dundee has a customised performance space situated within the Computer Science Department. This facilitated a series of innovative collaborations with Maggie Morgan, the founder of the Foxtrot Theatre Company to investigate the design of assistive technologies for older people.

An interdisciplinary approach to HCI research is not a new idea, and, at present, computer science researchers appear to be enthusiastically welcoming researchers and practitioners from unusual disciplines into collaborative partnerships.

The author is a member of the 'Leonardo.net':

"... an international, radically interdisciplinary research network led by the UK, set up to define a programme of research in the area of culture, creativity and interaction design. Drawing researchers from art, design, IT, computer science, engineering, architecture, cultural studies, and media studies, to name a few..." (Dix 2008).

And in the field of synthetic speech the author is a member of 'Netvotech':

a "Network on technology and healthy human voice in performance", which aims to "include practicing creative vocal artists and professional voice coaches, as well as scientists and engineers" (Netvotech 2008).

There are pitfalls to this approach. The criticism is that it creates an ethos of non-specialism, or of 'dabbling', as both groups attempt to come to grips with a 'foreign' discipline's specific languages and conventions. The interdisciplinary team may naively apply 'new' understandings to problems with long-established specialised methodologies and theories already in place, effectively re-inventing the wheel. This may sometimes occur, but not to attempt interdisciplinary research because there is a risk of failure seems contrary to the fundamental research ethos.

The intention of this thesis is to elicit support from the arts to solve the challenges scientists face in designing better synthetic speech production technologies. The two disciplines - science and art - employ radically different methodologies. It has been difficult to reconcile the scientific requirement for hard evidence and robust theories with the performing arts' prevalent suspicion of both; however, the reconciliation of these approaches may be vital in developing a truly holistic approach to the complex design challenges faced in synthetic speech production.

Acknowledgements

This research has been conducted part-time, and therefore has taken seven years to complete. I would like to thank my supervisor Alistair Edwards at the University of York for sticking with it with such enthusiasm for such a long time. I have relied on both his journalistic skills and attention to detail. Thanks to Paul Cairns at York for pointing me in the right direction with the statistical analysis. I have also been lucky to have had inspirational chats on all sorts of speech related topics with David Howard and Roger Moore at the universities of York and Sheffield respectively.

While conducting this research, I started employment at the University of Hull and I would like to thank Linda Hockley and Fiona Bannon for allowing me the freedom to complete the work on time, particularly in the write-up stage. I would also like to thank the University of Hull for the financial assistance I received over the years to support this research.

My thanks to colleagues and students who have participated in the studies and to the audience at the Post-Christmas Blues concert on the 19th of January, 2008, whose evening was interrupted by a 'user survey in pauses in synthetic speech'. Thanks also to the anonymous contributors to various web-based surveys and tests. The performers and audience at the 'PAT Testing' performance deserve a special mention for perseverance, after enduring an evening of performances of obscure modernist, period and newly commissioned texts by a synthetic voice.

I would like to thank Sir Alan Ayckbourn for allowing me to use the characters from his play 'Comic Potential' for a new scene. This scene was specially written for the 'PAT Testing' performances by Stephen Chance, together with whom I would like to extend thanks to all my creative collaborators on the exhibitions and installations documented in this thesis. They are Stuart Andrews, Stephen Chance, Chris Curtis, Paul Elsam, Ian Gibson, Andrew Head, Tim Howle, Julie MacDermott and Barry Perks.

From my pre-research life, I would like to thank Sir Peter Hall for teaching me just about everything I know about theatre and opera, and for being prepared to patiently mentor a novice director for five years. It was largely his influence, and that of the composer, and my dear friend, Paul Barker that inspired my interest in speech and singing.

All my immediate family have had put up with my obsession. This has included numerous holidays in which I would either be thinking or writing, as well as countless hours spent in my study in which I would be incommunicado. I thank all of them for never complaining and for always being so good humoured. My eldest son George has been a diligent proof-reader, and my younger son Arthur has restrained from loud drumming practice when I was working. My wife, the opera singer Maria Bovino, has been an important sounding board for ideas and has provided her beautiful voice whenever it was needed. What a shame she was not required to sing for any of the experiments. She has provided unflinching support and encouragement for my 'project', and, for this, she is truly a saint.

I am very sad that my mother did not live to see this work completed. So much of the creative side of this research is inspired by her love of the arts, particularly Shakespeare and opera. My father has been a model of personal perseverance, who, like me, discovered education later in life. His devotion to my mother and the rest of the family has been the bedrock for all of this.

Declaration

A combined total of eleven papers, conference presentations and exhibitions have been produced by Newell and Edwards, based on materials included in this thesis. These are listed in Appendix B. The artworks and installations developed to demonstrate or test findings are all collaborations, and the work of the artists who contributed to these projects is acknowledged as appropriate in this thesis.

I was assisted by Alistair Edwards at the University of York and James Proctor at the University of Hull in translating the code for filtered randomness (Rabin 2004) from C++ (which I don't know) to Microsoft Visual Basic (which I do). When programming the PAT software tool, I also made use of the public domain libraries in Leslie Sandford's MIDI Toolkit (Sandford 2007). The sequence of lip-synched animated mouth graphics used in the 'Not I (reprise)' performance in 'PAT Testing' were modified versions of the 'Preston Blair Phoneme series' which are available freely on the World Wide Web in many different formats.

I declare all the other work presented in this thesis as my own.

Christopher Newell 2009

To Maria

1 Introduction

*“In technological modernity, the dead and dumb word of matter begins to speak, though not now as the voice of nature or the breath of God, but on its own.”
(Connor 2000, p.42)*

In this chapter the broad concepts addressed by the thesis are briefly introduced. The title of the thesis and the motivation for the research are explained. The methodologies applied during the research are outlined. The research’s constraints are set out. The three key propositions are itemised and a summary of chapter contents is provided.

1.1.1 Discussion

Imagine you are through to a telephone call centre and encounter an automated digital message informing you that you have to wait to get through to a human operator. No effort has been made to disguise the fact that the message is being relayed by a machine; in fact, the unrealistic characteristics of the machine voice have been emphasised. It is also quite clear that the machine is reading from a script. Can this brazenly artificial experience be in some way ‘better’ than an encounter with a recording of a human speaker; and, if so, how? If it can, the content provided to the listener can be updated without employing voice actors, the script can be dynamically generated and processed using text-to-speech (TTS) technology, and multiple languages can be provided easily. If a non-realistic voice improves the experience, doesn’t this have serious implications for the current drive for more realistic, human-like synthetic voices? In turn, doesn’t this have broader implications for artificial intelligence and other disciplines that seek to synthesise anthropomorphic (human-like) features and behaviours for use in computer systems?

The subject of this thesis, when presented as a discussion, exposes concepts that some readers may find troublesome. The mythology surrounding artificial voices, from the Delphic oracle to

HAL (Kubrick & Clarke 1968), via séances, Von Kempelen's talking machine (Standage 2002) , Sparky's Magic Piano³ (Capitol 1947) and 'Fitter Happier' by Radiohead (Radiohead 1997), is not positive.



Figure 1: Von Kempelen's "synthesizer" (image from (Black 2009)), Sparky's Magic Piano (image from (Flickr 2009)) and HAL (Kubrick & Clarke 1968)

Artificial voices are typically portrayed as malevolent (HAL), bullying (The Magic Piano) or existential (Radiohead). Underlying this hostility is a deep unease with artificial voices, and a confidence in the uniqueness of the human spirit and its embodiment in the human voice.

"Not every sound made by an animal is voice (even with the tongue we may merely make a sound which is not voice, or without the tongue as in coughing); what produces the impact must have soul in it and must be accompanied by an act of imagination, for voice is sound with a meaning." (Connor 2000, p.24) citing Aristotle, *De Anima* 2.8.

The voice is regarded as more than a mode of human communication; it appears to embody the essence of a living person projecting their personality and even their soul to the outside world. Just as disembodied voices (voices with no bodily source), such as the voice of the Wizard in the Wizard of Oz (Baum, Garland, Haley et al. 1939, 1988), are perceived to be threatening or frightening as they deprive the listener of the anchor that locates the sound to its source, so artificial voices (voices that can have no bodily source) which nonetheless appear human are regarded as duplicitous; as implying the presence of a source that is not there. Should, therefore, a synthetic voice be presented so that it may be mistaken for a real voice, or is it preferable that a synthetic voice be always recognisably artificial, in the hopes that such transparency will side-step the listener's unease? This thesis attempts to address this

³ The voice of the 'Magic Piano' is produced using a 'vocoder', a technology that predates speech synthesis as it is understood today.

philosophical problem and to resolve it with a technical solution for text-to-speech synthesis systems.

1.1.2 Evolution of the research

This thesis arose from the researcher's theatrical production experience with human performers (see Appendix A). Anecdotal evidence from performances seems to suggest that voice actors and singers are able to encode complex emotional triggers in their auditory performances, based on features in the script or score but not on a semantic interpretation of the content of the text. These 'tricks' could be used to manipulate audience responses. Could these techniques be applied to synthetic speech without making more fundamental changes to the quality of the voice?

The techniques in question are exemplified by some classical actors who admit to 'playing Shakespeare', as one might technically play music: voicing the right sounds without understanding or necessarily wishing to understand what the words actually 'mean'⁴. In contrast, modern acting literature concentrates on psychological techniques (Strasberg & Morphos 1988, Stanislavskii 1948) that require a close alignment between an actor's psychological state and that of the character she plays as expressed by the text of the play. Certain techniques based at least partly on prosodic variations disassociated from meaning can be found in some specialized forms, such as voice acting, verse speaking and elocution (Hall 2004, Tucker 2002, Barton 1984, Alburger 2002). Among an arsenal of techniques available to the actor are simple prosodic variations such as the 'dramatic' pause. Dramatic pauses of indeterminate length can be used by actors to get the attention of the audience, to add weight and emphasis to a speech, or to disguise a temporary memory lapse. Anecdotally, it seems that the duration and position of a pause, within reasonable bounds, can be arbitrary without altering the pause's positive effect on the audience. Some of the studies reported in this thesis test this proposal on a synthesised speech stream. A question may be framed thus: can an algorithm for appealing speech sounds, based on prosodic/paralinguistic variations and concentrating on pauses, be encoded in a non-semantically derived form? Were an algorithmic

⁴ Conversations between the author and actors appearing in "The Late Shakespeare's" at the National Theatre, London, UK, May 1988. The production is documented in 'Peter Hall [on the Late Plays]' (1989) available from; 'Shakespeare in Performance' as a contribution to the 'Internet Shakespeare Editions' at <http://internetshakespeare.uvic.ca/Theater/sip/production/recorded/746/index.html> [accessed 18/08/09]

solution to be specified and proven to be sufficiently generalisable, could it be applicable to any speech stream in any context, including computer generated speech?

In addressing this question what emerges presents a more complex picture. Improved results in lab-based listening tests with extra embedded pauses were initially encouraging. The results from subsequent field tests during which users could interact with the speech system in a more realistic environment appeared to show that a positive result rested on a more complex relationship between *how* the voice spoke (regardless of extra pauses), *what* it said, and *where* it said it. This result was in effect a demonstration of the significance of the same ‘how, what and where’ variables applicable to real-world human speech and the negation of the ‘quick and dirty’ prosodic/paralinguistic variation proposed above. Of interest was that the studies showed ambivalence in the user response that did not exclusively rest on how the voice spoke (the focus of most research efforts in synthetic speech). Tests showed that how the voice spoke could be out of step with the other two variables and the voice could still be reported favourably. Could it be that a poor speech synthesiser with a good script in the right setting could succeed?

Thus an alternative research agenda arose: to demonstrate the value of a framework which describes a complex interrelational problem space⁵ in which a synthetic voice artefact⁶ can be presented. The result of applying the framework is to identify issues related to the problem space that can be carried forward to the design stage, preferably within an interdisciplinary design team.

The framework posits a re-focusing of priorities in the design of synthetic speech artefacts away from considerations of voice quality in isolation and towards considerations of voice quality as part of a multidimensional speech-production event.

1.1.3 What is the PAT framework?

This thesis describes a design framework. A framework can be thought of as “advice for designers as to what to design or look for.” (Sharp, Rogers & Preece 2007, p.84) For other

⁵ For an overview of the notions of ‘problem spaces’ and how they lead to improved conceptualization within the ‘design space’, see (Sharp, Rogers & Preece 2007, pp.46-50).

⁶ The term artefact is used throughout this thesis to mean more than just a synthesis system. It implies a comprehensive speech embodiment that may include hardware and software as well as accoutrements associated with art and performance such as a set and a script.

frameworks applicable to HCI in general, see (Carroll 2003). The PAT framework is developed, as specified by Sharp et al., based on “the experiences of actual design practice and findings arising from user studies” (Sharp, Rogers & Preece *op. cit.*) The PAT framework is a high-level framework designed to support the presentation of the ‘problem space’ (*ibid.* p. 46) within a set of useful criteria (dimensions) that map to potential design solutions. By ‘high-level’, it is meant that the framework deals with the voice at the output or performance⁷ stage when speech sounds are perceived and interpreted by the user. To describe it as ‘low-level’ would imply considerations of computer linguistics, natural language processing and algorithms for text-to-speech synthesis. These factors are not addressed in this thesis. At present there is no other framework offered to designers of synthetic voices focused on the performance stage. The reason for this may be that such considerations can only come to be regarded as significant once the low level issues of speech production (how the voice speaks) have been, at least partially, solved. It may also be the case that such issues are perceived as tasks for somebody other than a synthetic speech engineer or speech researcher to address. Accordingly, in the studies documented in this thesis a collaborative interdisciplinary design approach is taken, with contributions from script writers and other art and performance practitioners. Thus the framework can be seen as a high-level, multi-dimensional and multi-disciplinary perspective on the implementation of synthetic voice systems. The intention is not to develop a new synthetic voice or solve low level technical issues; rather, it is to develop a framework that improves the user evaluation of new and existing synthetic voices. The viewpoint is that of a user or audience scrutinising the ‘performance’ (in the theatrical sense) delivered by a synthetic voice ‘actor’. The question posed is: can the application of the PAT framework improve the appeal (through liveness) of synthetic speech reported by such a user? An illustration of the context for the framework is shown in Figure 2.

⁷‘Performance’ as in ‘theatrical performance’ not as in ‘improving performance.’

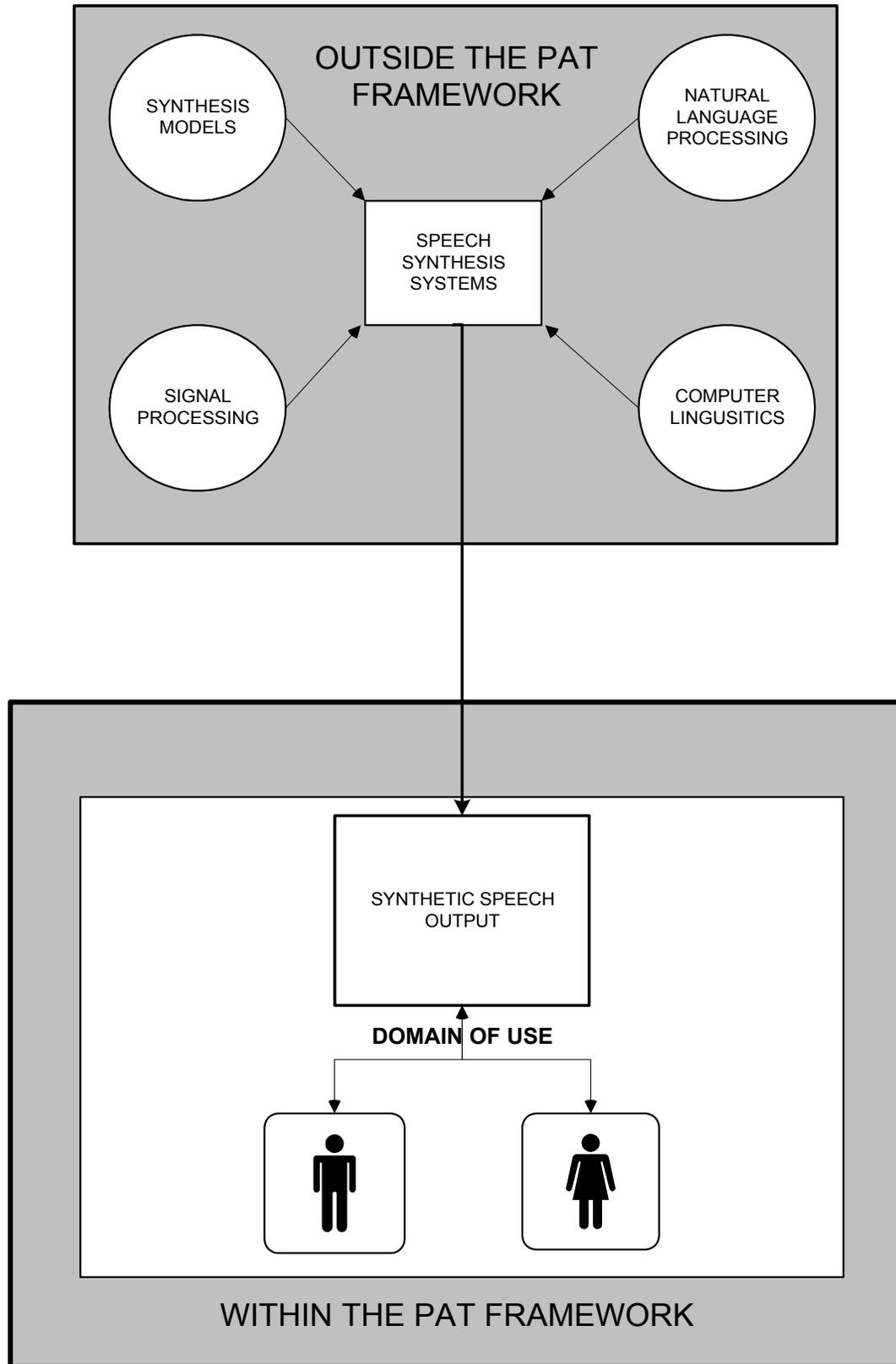


Figure 2: The context for the PAT framework

1.1.4 Text-to-speech synthesis and the framework

The technology most likely to benefit from the framework - also used to test the framework's validity - is text-to-speech synthesis (TTS). TTS outputs computer-synthesised speech audio from text input such as word processor documents. The synthetic voice will speak the words as clearly and naturally⁸ as the technology permits. These days the user of a desktop system will have access to a choice of TTS voices which may offer different genders, languages and speech styles depending on the operating system and any bespoke speech software installed. In addition to desktop applications, TTS is implemented in accessibility systems, web based systems, mobile devices, gaming and telephony. Narayanan & Alwan provide an overview of recent developments in TTS (Narayanan & Alwan 2005).

The principal advantage of using TTS rather than pre-recorded or rendered speech chunks is that a TTS system should be able to speak any text input, including arbitrary text (for example, an RSS feed or user-entered text) giving the illusion of real-time⁹ (without a perceptible delay). However, this is at the cost of voice refinement and accuracy. Some modern synthetic speech generation systems offer editing tools to manipulate the speech production parameters, e.g. (Loquendo 2008, Cepstral 2006) to produce more realistic or specific effects (a Loquendo example is included on CD Track 5). However, because of the generic requirements a TTS system has for processing arbitrary text, such refinements are not wholly viable. Currently TTS systems are prone to errors (mainly in parsing, pronunciation and prosody); thus, directing the user to strategies for accepting the inevitable errors is part of the PAT framework's purpose.

The framework is applicable to any synthetic voice implementation, but may offer fewer advantages in those domains where the processing of arbitrary texts is not necessary. For example some operating environments require only a pre-determined, limited set of utterances, subject to only occasional content updates; for example, a game, simple expert system, or telephone call routing system. In these circumstances, TTS may not be appropriate and the choices facing the designers may lie in selecting from the available pre-rendering synthesis systems or using recordings of human actors. In many instances, a human recording

⁸ Comparative terms such as 'naturally' are discussed in Chapter 3 The language of speech related phenomena.

⁹ Real time [in this thesis] refers to sensing and responding to external events nearly simultaneously (e.g. within milliseconds or microseconds). From 'The Linux Information Project' at: http://www.linfo.org/real_time.html [accessed: 19 August 2009]

may be the best choice. The PAT framework is ideally suited to implementations where the advantages of dynamic or arbitrary text-based content generation and processing outweigh the disadvantages of relatively 'odd-sounding' speech. Examples could be a web2 system with user-based content generation, a screen reader, or a dynamic news reading service. However, the framework can provide insights into the manipulation of the user acceptance of any synthetic speech system, TTS or otherwise.

1.1.5 The current state of speech synthesis

HCI research recognises the potential value of speech-based interfaces (Weinschenk & Barker 2000, Edwards 1991, Nass & Brave 2005, Ballentine 2007). Not only is speech the natural way humans interact with other humans, but speech also provides the only way for some users with specific needs, principally blind people, to interact with computers (Edwards *ibid.*). At first, synthetic speech production was considered a trivial task and, in 'The Age of Intelligent Machines', (Kurzweil 1990) anticipates comprehensive speech capabilities for machines to have been developed by now. Although we are some way off the kind of effortless speech interaction and seamless anthropomorphism predicted by Kurzweil and vividly dramatised by HAL in '2001: A Space Odyssey' (Kubrick & Clarke *op. cit.*), reasonably clear, precise synthetic speech production has been achieved in a number of languages, such that Keller can claim that "Some systems are so good that one even wonders if the recording was authentic or manufactured." Preface to (Keller 2002). According to Black the intelligibility of some synthesis systems easily exceeds that of human speech particularly if exposure to the synthetic speech is repeated (Black 2009).

This perspective may appear dangerously complacent, but from such a perspective we may usefully reflect on the high quality of the multiplicity of voices that have been produced so far and consider ways in which their 'performances' can be improved. However, this is not the prevalent perspective and the success in producing high levels of intelligibility has led developers to evolve new, more challenging, requirements, arguably missing the opportunity to implement potential 'quick and dirty' techniques, such as the proposal herein, for improving what they have already produced. According to these new requirements it is no longer sufficient for a synthetic voice to speak reasonably clearly and be understood, they may also be required to exhibit personality, mood, character and a range of other complex human

behaviours and characteristics. These are problematic terms with no agreed definitions across the contributory disciplines - an issue which is addressed in Chapter 3. However, there is agreement that the automated parametisation and control of the auditory components that can encode these features is highly individual and complex; hence the current synthesis solutions tend to be focused around the manipulation of large databases referencing human-sourced samples (unit selection or concatenative synthesis (CD Track 3)). By using a human source, the systems are designed to retain some of the complex characteristics referred to above, as well as to be more pleasant to the ear. Different voice styles can be encoded by using different source voices or versatile actors. Difficulties arise due to the amount of data required, and the time needed to digitally record the actors' voices - particularly if the intention is to exhibit the range of voice styles listeners expect (Keller 2002, p.10).

Formant synthesis (waveform construction from formant models) (CD Track 1) developed in the 1960s, is less data-intensive technology than unit selection or concatenative synthesis and probably the sound that most people associate with synthetic speech. It tends to produce the machine-like quality that characterises the sound made by Professor Stephen Hawking's synthesis system and some of the classic robotic voices popularised in film and other media. Diphone synthesis (CD Track 2) developed in the 1980s, is the precursor of the unit selection synthesis described above. It selectively concatenates a small number of instances of speech. Statistical parametric synthesis (Black, Zen & Tokuda 2007) constructs waveforms from parametric models and is at the cutting edge of current research¹⁰. Articulatory synthesis (CD Track 4), such as the one packaged with the PRAAT software package (Boersma & Weenink 2008) relies on the mathematical modeling of the human vocal tract and becomes very close to human-sounding speech under ideal conditions. However, ideal conditions rarely present themselves in real world TTS implementations, and, as none of the current synthesis models reliably produce natural human-sounding speech, the user acceptance (or enthusiasm for) synthetic speech, even in specialised application domains, appears to be a significant HCI challenge¹¹ (Ballentine *op. cit.*).

¹⁰ Further examples of statistical parametric, unit selection and diphone synthesis can be found at the 'Festival Online Demo.' Found at: <http://www.cstr.ed.ac.uk/projects/festival/morevoices.html>. [Accessed 22/08/09]

¹¹ I am indebted to Professor Alan Black from Carnegie Mellon University for use of his presentation notes when preparing this brief historical overview.

Moving briefly into a specialised application domain relevant to this research, the take-up for applications of synthetic speech in an artistic domain has been very limited. In the early 1970s the composer Charles Dodge was given the opportunity to work with a pioneering voice synthesiser under development at Bell Labs in New Jersey. The ‘Speech Songs’ he produced were the first example of ‘Art’ produced using a speech synthesis system. Since then there have been a few other examples most notably Radiohead’s song ‘Fitter Happier’ (Radiohead *op. cit.*) and Laurie Anderson¹² for ‘Oh Superman’ (Anderson 1981). Extracts from several audio artworks utilising speech synthesis are included on the CD Track 6 to Track 11. The modest uptake may be due to the technical knowledge required to manipulate the low-level properties of the speech stream. In 2004, with the advent of Synthetic Speech Mark up Language (SSML), this has become easier.

Despite the prevalent view that speech is the most intuitive way for man and machine to communicate, there is no uniform acceptance that human-like speech-based interaction is beneficial to HCI. Nass & Brave suggest that human-like synthetic speech facilitates improvements in HCI by eliciting the same interactive responses from human subjects as normal human-to-human speech interaction (Nass & Brave *op. cit.*). Ballentine cites many examples where speech interaction is less successful than conventional keyboard and mouse interaction, and is broadly opposed to the blanket assumption that speech-based interaction is best (Ballentine *op. cit.*). Moving on to consider the broader context of human-like machines, Mori writing in the context of robotics, suggests that there is a limit to the degree of anthropomorphism that can be applied to a non-human embodiment. He calls this the ‘The Uncanny Valley’ (Mori 1970). A non-realistic entity with limited human traits may actually be more acceptable to users than examples that are more fully anthropomorphised - those which, for example, appear to breathe, have human-like skin or exhibit human social behaviour. These enhancements, designed to make the artefact more ‘normal’, may have the reverse effect and appear ‘uncanny’, alienating the user. Figure 3 shows an example of a modern robot some may find ‘uncanny.’ Mori’s theory is discussed in more detail in section 2.3.

¹² Strictly speaking Laurie Anderson made use of the ‘Vocoder’ for ‘Oh Superman’ which as previously noted in the case of ‘Sparky’s Magic Piano’ is not speech synthesis technology.



Figure 3: Professor Hiroshi Ishiguro (standing) and Repliee Q1Expo (sitting)¹³

Underlying the debate for and against synthetic speech appears to be a craving for normalisation or realism. The assumption is that, by ‘normalising’ the abnormalities inherent in synthetic speech, HCI will improve. It may be argued that aiming for a synthetic voice that is ‘normal’ may be an unrealisable objective. We do not expect all human voices to be normal. We accept heavily accented speech, individual nuance, dysfluencies and disorders. We expect personal and ‘inappropriate’ variation at all levels from the prosodic and paralinguistic to the lexical, and yet we are encouraged to expect a synthetic voice to show only ‘appropriate’ variation. It may be that synthetic voices will always have a ‘computer’ accent, make computer-like errors and will choose words that are appropriate to a computer, rather than a human. As Alan Turing says, “It might be argued that when playing the ‘imitation game’ the best strategy for the machine may possibly be something other than the imitation of the behavior of a man” (Turing 1950). Donald Norman makes a similar point when discussing robots and emotions: “The affect that robots require will be both similar to and very different from that of people” (Norman 2004, p.167). It may be that the normalisation and drive for realistic voices is not achievable and that voices that are both similar to and very different from people are required for computers.

In the next section, a broad context for the meaning and evaluation of realism is presented.

¹³ Image found at <http://www.is.sys.es.osaka-u.ac.jp/index.en.html>

1.1.6 Realism v the willing suspension of disbelief

Representing realism is crucial in the design of many different types of artefacts, from theme park rides to digital pianos. Computer games designers have been particularly keen to exploit the public appetite for realism with increasingly powerful graphics cards providing ever more-realistic visuals. Animated films have striven for very realistic environments and in some cases representations of very realistic humans. For an example, see the film 'Final Fantasy: The Spirits Within' (Sakaguchi, Lee, Baldwin et al. 2001). High-definition video and high-fidelity audio are standard in many Western households, and computer-generated effects in film are increasingly realistic, often making it impossible to determine what was actually filmed and what was subsequently superimposed on the film during 'post-production'. In computer games, notions of reality are addressed both in terms of the visual and auditory representation of the game-world and its inhabitants and also in terms of the 'realistic' interactive experience offered to the player. These examples take an essentially literal approach to the notion of representing realism, assuming that greater verisimilitude to the human experience of reality by adding detail (increases in the resolution) increases the perception of reality for the recipient. Of course, as the resolution approaches the limits of human perception, the improvements become smaller and the rewards less worthwhile.

Theatre has shown that the representation of realism can take many forms. Visual realism as presented above had a brief and largely unsuccessful vogue in theatre set design and was subsequently abandoned, but the quest for realism continues to define the predominant acting style in TV and cinema today. In theatre, 'realistic acting' concerns the underlying truthfulness of the actor's performance. The actor is required to believe the situation they are to represent is true in order to present a credible performance. Clearly this is a problematic concept for speech synthesis. The term 'naturalism' as defined in theatre presents a more computationally viable option: "At its most basic, it is a facsimile, a replica of the surface, nothing more" (Benedetti 2007, p.102). But in most instances in theatre a basic facsimile is still not viable and such verisimilitude is impractical, and thus levels of abstraction are applied for both practical and aesthetic reasons. The aesthetic argument is very potent and has given rise to representations of reality with significantly reduced resource overheads. Nowadays, for example, it is rare to see a stage-set with real doors and windows if lighting or projections can adequately represent them. A phenomenon occurs here which was described by Coleridge as

'The willing suspension of disbelief' (Coleridge 1817) (for brevity's sake henceforth referred to as WSOD). A user absorbed in a computer game, listening to a radio play or even reading a book agrees to accept the illusion evoked by the game designer, actors or author, and, for the duration of the experience, to suspend disbelief rather than crave realism. The effectiveness of the conceit is mutually negotiated and is different for different authors and for different users. WSOD shares some features with a phenomenon identified by Csikszentmihalyi as 'flow' (Csikszentmihalyi 1990) in which the immersive nature of the experience causes the user to lose consciousness of the artificiality of the event. In WSOD, however, the user is always aware of this artificiality, but chooses to repress consciousness. In either case, the event must offer emotional 'rewards' – excitement, escapism, amusement - that encourage the user to submit to the experience. As Hayes-Roth says "... animate characters will pass our test of life-likeness, when people willingly suspend disbelief in order to have the service, pleasure, convenience or other desired consequences of interacting with them as if they were real" (Hayes-Roth 2003). Applying WSOD to this research, a voice that makes no concessions to realism must offer some other kind of 'reward' in order for the user to be prepared to suspend her disbelief. One suggestion for a reward is the perception of 'liveness,' the subject of the next section.

1.1.7 Liveness

'Liveness' (Auslander 1999) is not a term in common usage, nor one that is easy to define. Auslander sets out a complex cultural and political discourse that draws from a wide range of literary, performance and media theories that have been comprehensively reviewed in the performance literature and are outside the scope of this thesis. As such 'liveness' is a conflation of many concepts, and has no single, fixed, definition beyond a vague sense of immediacy and 'now-ness'. It can be described as the difference between speaking and reading, between the spontaneous and the scripted; but such a description overlooks the ambiguous territory of the 'mediatised' live performance – defined by Auslander as "performance that is circulated on television, as audio or video recordings, and in other forms based on the technologies of reproduction" (ibid p.5). It is plain that Auslander does not make a binary distinction between the performed and the 'real' (or non-performed); and it is this grey area that this research will attempt to exploit. A synthetic voice is doubly mediatised, as both the production and the means of reproduction are exclusively technologically-based. Subtleties emerge when we consider the types of synthesis systems, in some cases the

resolutions of the sampling technologies employed to capture and store the original human sources. However, Auslander believes that: "...all performances live or mediated are now equal" (ibid p.50) - equal in the sense of the user's broad acceptance of them. From this we may extrapolate that, for the user, a synthetic voice can be perceived to possess equal or even greater 'liveness' than a human voice if certain conditions are met.

This research does not attempt to critically review liveness, but rather to present a codification of liveness applicable to a computer-based rendering. It follows a breadcrumb trail of auditory clues, revealed by Auslander's approach, that a computer might be able to process and represent for consumption by the user. It will be no surprise, given that an artificially intelligent solution is not contemplated, that some significant quantisation has to take place and the analogue nuance of the compressed version of Auslander's theory we have adopted has to be sacrificed to the binary determinism of a calculating machine. For now, the reader may also have to accept that the notion of digital 'liveness' is not oxymoronic.

It may be helpful to state that in this research, 'liveness' is not, as Ednie-Brown says, 'The sense of a living presence' (Ednie-Brown 2004) although that 'sense' remains a useful shorthand for explaining 'liveness' to an audience unfamiliar with the theory. In order to narrow the scope, and to focus on features that permit machine rendering, we have codified liveness as a simple continuum by which an audience (users) can plot the degree to which a performance event is happening, or has happened, live. At one extreme on the continuum is a live event which a live audience experiences in real-time, while at the other extreme is an event which has been recorded and is observed by the audience at a different time and in a different place to that of the original.

Examples might be the difference between watching a live performance by a string quartet in a concert hall (liveness at a maximum) and listening to a CD of a studio recording of the same quartet while travelling in the car (liveness at a minimum). Precise mapping of an event on the continuum is not the point. Rather, it is identifying features that signify to the user one position relative to another that may be of interest. Referring again to the string quartet example, it may be audible features such as the sound of page turns (normally suppressed by performers in studio recordings), mistakes in the performance, or just the difference in the acoustic environments (e.g. the amount of reverberation) that signal the degree of liveness to the listener. On first reading, 'liveness' appears to be readily perceivable, offering little granularity;

however, in 2.5.9 the theory will come under greater scrutiny and degrees of liveness will be revealed.

CD Track 12 and Track 13 provide a particularly vivid comparative example. Track 13 is a recording of a scene from an opera recorded in a recording studio. Track 12 is the same scene with the same singer recorded in concert in Parma, Italy. Despite the poor-quality recording, the concert recording seems to have greater liveness. Even greater liveness would have necessitated being present in the audience when the recording was made.

To summarise: synthesising liveness in the PAT framework involves a combination of techniques designed to manipulate the user's perception of where the speech event lies on the continuum set out above. These techniques aim to evoke 'liveness' for the user without necessarily projecting a realistic living presence¹⁴.

1.1.8 Place, Authenticity and Time

As we have discussed, human actors, dramatists and musicians have an arsenal of sophisticated techniques that they employ to manipulate WSOD and simulate liveness. Examples of such techniques include an author framing a novel as a collection of letters or diary found by chance or informing the reader that the story is based on 'real' events when it is not, or a record producer inserting remnants of studio chat at the end of a recording or adding guitar fret noise, neither being part of the original recording. We will posit that similar liveness-simulating modifications to a synthetic speech artefact can be plotted on three dimensions:

1. *Place*: locating the voice and the user in a shared auditory and physical space (be it fictional or actual).
2. *Authenticity*: manipulating the user's sense that the voice is being transparent or truthful.
3. *Time*: manipulating the user's perception that they are operating in a time frame shared with the voice.

'Place, Authenticity and Time' yields the acronym PAT. Each dimension can be rendered using conventional theatrical combinations of scripting, setting and acting, as might be applied to a

¹⁴ If the clumsiness of the term 'liveness' becomes burdensome to the reader the term 'spontaneity' may be an appropriate temporary substitute, although it fails to capture the multidimensional qualities imbued in liveness.

radio-play or other text-based theatrical presentation. The script is the actual words spoken by the voice, provided by the script writer. The setting is the location of the voice agreed with the user (this can be fictional or actual). The acting is the voice itself, and could include all the modifications available to the specific voice; modifications which might include voice style, gender, voice quality, and all paralinguistic and prosodic variables. These are similar to the modifications a human actor has available. In this research we have limited acting to a small set of prosodic variations that can be realistically tested.

Dimension	Render method	Outcome
Place = location of voice + location of interface	Script	Liveness _p is optimised either when the voice's place and the user's place are perceived as the same, or when a place is agreed with the user.
	Setting	
	Acting	
Authenticity = perception of truthfulness +/- suspension of disbelief rating	Script	Liveness _A is optimised when the perception of truthfulness is highest or when belief is suspended the most completely by the user.
	Setting	
	Acting	
Time = real-time +/- agreed time	Script	Liveness _T is optimised when the voice is perceived as operating in real-time, or at a time that is agreed on with the user.
	Setting	
	Acting	

Table 1: Table showing the Place, Authenticity, Time (PAT) framework

Table 1 shows a tabular version of the PAT framework in which the three dimensions of liveness are rendered using script, setting and acting. The optimal outcome for liveness_p, liveness_A, liveness_T, is described.

1.1.9 Testing the PAT framework

The PAT framework is designed to help synthetic speech designers review design decisions within a broader problem space that could be characterised as 'beyond realism' or unrealistic. Paradoxically, PAT is an attempt to position a synthetic voice within a '*realistic*' framework of user perceptions and expectations and implementation-specific requirements that do not

speculate on wide user acceptance of a voice that sounds nearly, but not quite human. There would be little point in proposing this unless it can be shown to work.

The framework will be required to posit technological methods of realising the theoretical manipulations based on the PAT dimensions. These methods will be realised through script, setting and acting using existing synthetic voices: this is done using the PAT software tool. In particular the software is designed to demonstrate that modest prosodic/paralinguistic changes, pauses and speech rate variations made to an existing synthetic voice will improve user evaluations and results in instances of perceived liveness. The metrics applied to the prosodic variables modified by the PAT software tool are based on the sources described in Chapter 2 Sources, perspectives and features for the PAT framework. The tested voice is 'Microsoft Mary', one of a number of voices provided with the Microsoft ® operating system. The Microsoft TTS system is compatible with Synthetic Speech Mark-Up Language (SSML) (W3C 2005), and facilitates modifications to the audio output of 'Microsoft Mary' in something like real-time (with no appreciable delay). In addition, it is possible to present the modifications in a variety of experimental and more realistic environments designed to explore the framework dimensions. These included the web, telephony and live theatre audiences. The results from these studies and performances provide the evidence from which conclusions are drawn.

1.1.10 The propositions examined in this thesis

This thesis will examine the following three propositions:

- **The first proposition is that the automated modification of pauses and speech rate variations can be shown to improve the user perception of liveness of a synthetic voice when compared to the same voice that has not been subject to such modifications.**
- **The second proposition is that the PAT framework provides a conceptual model within which to negotiate the problem space presented to designers of synthetic voices in an interdisciplinary environment.**
- **The third proposition is that the PAT framework can be shown to improve the user perception of liveness in a synthetic voice artefact when compared to the same voice that has not been subject to the modifications rendered by the PAT framework.**

1.1.11 Defining sources, perspectives and features

The differences between the sources and the perspectives discussed in this thesis that present potential insights into the development of a broad understanding of the human/synthetic voice relationship and those that provide codifiable features are complex. Some sources of knowledge or information, such as Renaissance verse speaking, provide both; others such as the 'acousmètre' (Chion & Gorbman 1999) do not present codifiable features. To help, the terms presented in Table 2 are used consistently throughout the thesis to describe the type of knowledge or information under discussion and the degree of precision or codification that knowledge or information provides.

Highest level (the least precise)		
Term	Explanation	Examples
Sources	Philosophical, scientific or culturally determined positions from which complex systems of knowledge can be described and understood. The position will determine a generally agreed usage of a common vocabulary although some variability will exist	<ul style="list-style-type: none"> • Artificial intelligence • Performance • Music
Perspectives	A subset of sources defining a specific theoretical viewpoint that may have less general agreement in the field and will set out a vocabulary that may be less incontrovertible	<ul style="list-style-type: none"> • ‘Liveness’ • Renaissance verse speaking • Musical expression
Features	Specific components of perspectives or sources that can potentially give rise to coordinates, metrics or values	<ul style="list-style-type: none"> • The authenticity dimension • Iambic pentameters • Italian musical expression markings
Metrics	Numbers that can be used programmatically	<ul style="list-style-type: none"> • 22.6% of breath pauses in readings • 500ms

Lowest level (the most precise)

Table 2: Defining sources, perspectives, features and metrics

1.1.12 Methodologies applied in this research

The PAT framework emerged over the course of the research. The original intention was to concentrate on auditory modifications to a synthetic voice. The expectation was that sources from the performing arts would provide perspectives from which features could be codified as metrics and applied to the audio output of synthetic voices. These modifications could be tested against the original unmodified versions and conclusions drawn. It was hoped to find a generalisable algorithm that could be applied to any speech and any synthetic voice. The framework emerged when tests identified the significance of factors that could not be rendered as modifications to the speech signal and required a multidimensional mixed media

treatment. This led to the use of some unusual methodologies including user tests presented as theatrical performances and art installations as well as more conventional scientific studies. Table 3 charts the evolution of the framework over the duration of the research and the methodologies applied at each stage.

Commencement of research

Source	Methodology	Outcome	Information type	Example
HCI, Speech synthesis, Linguistics, Performing arts, Performance theory, Music	Literature Review	Potential new perspectives on the problem space	Knowledge	The distractive perspective as perceived in ventriloquism
English vocabularies and dictionaries	User questionnaire	Partial clarity and consensus on applicable vocabularies	Vocabularies, definitions	An agreed set of meanings for 'character voice'
Audio samples of human and processed speech	Lab based comparative test	User detection rate of auditory features	Statistical	'Liveliness' can be detected in content free speech-like sounds by (n) users
HCI, Speech synthesis, Linguistics, Performing arts, Performance theory, Music	Literature review	Potential for encoding human auditory features as synthetic speech	Auditory feature descriptors and metrics	The significance of grammatical pauses
Online synthetic speech application	Web based speech construction tool	User preference for specific speech features	Statistical	Preference for breath-like pauses (pauses at breath-like intervals) in (n) users
Exhibition of speech based artefacts	Demonstration and informal user review	Issues of user credence in background sounds	Anecdotal and qualitative	Breaths identified as uncanny
Synthetic speech production editing tool	Focus group	Refinement of speech production modification algorithms	Metrics	Longest legal grammatical pause set to (n) milliseconds (ms)
Installation of telephonic artefact	Field tests	Data contradicts previous studies. Significance of other dimensions emerges	Statistical	Extended grammatical pauses not preferred by (n) users in some settings.
HCI, Speech synthesis, Linguistics, Performing arts, Performance theory, Music	Literature re-review	Re-consideration of significance of new perspectives on the problem space	Knowledge	Significance of setting and scripting re-assessed
Presentation of multi-dimensional theatrical event	Audience survey	Complex multi-dimensional framework perceived	Statistical	User evaluation of voice speaking humorous text indicates increased perception of liveness by (n) users.

Submission of thesis

Table 3: Evolution of the PAT framework and the methodologies applied

1.1.13 Constraints applied to this research

In this section the constraints that can be said to apply to this research are set-out. This provides a context in which the limitations of this research's scope can be understood by the reader at the outset.

Speech production. This research is constrained to exploring synthetic speech production. Speech recognition is not addressed. User interaction with the speech output is constrained to the level of passive listening. However, by advocating the reduced use of human voice actors to produce the voices for synthetic speech systems, it may be easier to generate dynamic content for a synthetic speech system and thus improve the user experience and the potential for an enhanced interactive experience.

Technology. The technology employed for testing purposes is limited to the Microsoft Speech Application Programming Interface (Microsoft SAPI) and the only voice rigorously tested is 'Microsoft Mary' (see section 6.1.3 for the Microsoft voice specification). This is a constraint, and further research would need to be undertaken to test the rendering of specific acting modifications on other voices before any generalisable conclusions with relation to acting could be confirmed. The conclusions drawn with relation to acting are further constrained by the limit on the prosodic/paralinguistic variables chosen for manipulation. The choice to limit the variables to pauses and speech rate variations was deliberate and informed by the literature. The failure to show a positive user response in the field tests may have been the result of this constraint or it may have been related to the specifics of the voice rendering technology.

Biases in the evidence base. Ideally, at least from the perspectives set out in this research, synthetic speech should be improved by making modifications at the speech audio production and perception level alone. If it were possible to address this level in isolation then there would be no need for this framework and this research would be focused entirely on modifying the audio streams of computer-generated speech to increase user acceptance and more effectively evoke liveness. During the research the studies revealed the importance of other dimensions that required alternative rendering. Thus, of the three render methods described in the framework - acting, scripting and setting - only acting can be rendered entirely at the auditory level, by changing how the voice speaks; and, due to initially encouraging results, the majority

of the data gathered and the studies conducted relate to rendering acting. Thus the evidence base for the framework is biased toward auditory acting. This difficulty is made clearer in Chapter 7 'Evaluating the PAT framework.

Selecting texts for synthesis. The framework encompasses an infinite range of speech content modifications (texts) and settings. The choices made in these two areas for the studies were largely serendipitous although there is deliberate emphasis on telephonic settings and content largely because of the potential for real world implementation. With hindsight a more coherent methodology for the selection of texts for test purposes would have been beneficial. The literature revealed no consistent extant methodology for the choice of texts other than for the purpose of ensuring that the widest range of 'phones' (smallest audible speech segment) are rendered (Lea 1974): a requirement not relevant to this research. In the field the texts seemed to have been chosen largely at the whim of the researchers. This is illustrated by a fascinating listing at the Transcription of Recordings from the Smithsonian Speech Synthesis History Project (Smithsonian 2007). An effort needs to be made to rectify this issue with a standardised set of scripts (see Further research 9.6.8). The initial intent was to demonstrate a generic auditory solution that could be applied to any text therefore the choice of script was based on offering a large range of editing opportunities. As the framework emerged, and the significance of the script became apparent, the choice of texts became more precise.

Guidance not answers. The framework is intended to provide a perspective upon which the problem space facing the designer of a specific synthetic speech implementation may be more comprehensively and holistically assessed. The designer has to translate the broad principles set out in the framework to the specifics of the project and there is nothing in the framework to support this process. For example: knowing that a specific implementation may require 'liveness' expressed in the simulation of real-time hesitations is one thing; knowing how many hesitations fall in a specific speech segment still presents an aesthetic discretionary opportunity. The framework is constrained to provide guidance without presenting definitive answers.

Other minor technical constraints are described in the relevant sections of the thesis.

1.2 Chapter Summary

This section provides a brief summary of the contents of each chapter

1.2.1 Chapter 1: Introduction

In this chapter the broad concepts addressed by the thesis are briefly introduced. The title of the thesis and the motivation for the research are explained. The methodologies applied during the research are outlined. The research's constraints are set out. The three key propositions are itemised and a summary of chapter contents is provided.

1.2.2 Chapter 2: Sources, perspectives and features for the PAT framework

In this chapter, the sources and perspectives for the PAT framework are outlined. The challenge of user evaluation of synthetic speech is positioned in the context of a number of other fields that attempt to understand or represent varied aspects of human speech. Each source or perspective can be categorised as one of three classes: additive, distractive or subtractive. Some perspectives resist a reductionist approach, offering something akin to a framework for this framework, while others posit usable metrics. The perspectives include robotics, linguistics, acting and performance, music, speech in the cinema, and ventriloquism. Silence and negative spaces are significant contributors to a number of perspectives and an analysis of these phenomena will contribute to several sections in this chapter, with a discussion at the end. Finally, based on the sources and perspectives discussed, a theoretical basis for the PAT framework is presented.

1.2.3 Chapter 3: The language of speech related phenomena

This chapter sets out to define a structured vocabulary, derived from the relevant sources and perspectives discussed in Chapter 2, to provide a clear interpretation of the key comparative terms used in this document. The PAT framework circumvents the use of terms such as 'life-like', 'natural' and 'realistic', preferring a non-anthropomorphic approach. However, these broad concepts provide comparative benchmarks for the evaluation of auditory properties exemplified by extant synthesis solutions and require analysis. In addition, a short discussion is

presented exploring the difference between speaking and reading as it may relate to synthetic speech.

1.2.4 Chapter 4: Detecting and describing liveliness and liveness

This chapter examines the problem of defining ‘liveliness’. A number of different terms were considered for the dependent variable prior to the eventual selection of ‘liveness.’ The term liveliness was used in the first two studies reported in this research. To avoid confusion with subsequent tests in which other terms are used, these two studies are given a short chapter of their own. This chapter documents the two studies.

1.2.5 Chapter 5: Automated paralinguistic/prosodic modification

In order to test the validity of the synthetic speech the paralinguistic/prosodic modifiers identified in the review of sources and perspectives in chapter 2, and to meet the requirements of proposition one, each feature has to be rendered automatically in an appropriate synthesis system. This must be done both in isolation, and in combination with other features. A methodology for defining the heuristics and metrics for each feature has to be derived and a system, with a generalisable synthetic voice implementation, built to host the tests. The PAT software tool facilitates automated paralinguistic/prosodic manipulations to a SSML (*Synthetic Speech Mark-up Language*) compatible speech synthesiser. The input data is ‘plain text’ and the resulting limitations on the possible automated prosodic/paralinguistic modifications are set out in this chapter. The sources and perspectives determining the heuristics and metrics applied to the paralinguistic/prosodic modifiers implemented in the PAT software tool, briefly reviewed in chapter 2, are explained in greater detail.

1.2.6 Chapter 6: Evaluating the values for the paralinguistic/prosodic modifiers

This chapter documents the process of evaluating in a series of user studies the settings for the paralinguistic/prosodic modifiers implemented in the PAT software tool. Five studies were undertaken: a focus group study, three web-based pilot studies and a paper-based user survey. Two of the web-based pilot studies failed to produce reliable results. The other studies broadly supported the evidence from the literature, but the results directed the researcher to

consideration of a broader context for the successful rendering of ‘liveness’; a context subsequently developed into the PAT framework.

1.2.7 Chapter 7: Evaluating the PAT framework

The framework evolved as tests of manipulations made to the paralinguistic/prosodic modifiers in a synthetic speech stream using the PAT software tool indicated weaknesses in the initial assumptions and the potential significance of other variables on other dimensions became apparent. These alternative variables and dimensions are represented in the PAT framework. The evaluations are incorporated into an art exhibition, two telephonic art-and-performance works, and a theatrical performance. Methods are evolved to collect data assessing more complex levels of interactive engagement than those used in previous studies. The results reveal interesting but ambiguous trends in the user evaluation of synthetic speech artefacts.

1.2.8 Chapter 8: Potential applications

In this chapter, three scenarios are envisaged in which an interdisciplinary design team uses the PAT framework to explore the problem-space for a specific TTS implementation. The first scenario is a telephone banking system. The second scenario is for a portable exhibit guide system for a technology museum. The third scenario is for a toy polar bear teaching aid. In all cases the assumption is that the decision has been made to use TTS technology rather than samples of recorded human voices, at least for parts of the system implementation.

1.2.9 Chapter 9: Conclusion and future research

Chapter 9 draws together the results of the research relates them to the three propositions set out in the introduction and provides the final conclusions. A short section considering future research concludes the thesis.

1.3 Conclusion

In this chapter the main themes and technical challenges to be developed in the thesis have been presented together with a description of some of the constraints and difficulties experienced during the course of the studies undertaken. The essential dichotomy that exists between 'WSOD' and the 'Uncanny Valley' and the potential of a solution resting on the evocation of liveness will be the focus of this research. In addition this research addresses technical issues found in text-to-speech synthesis as well as non-technical issues found in performance and theatre and tries to derive solutions for the technical problems from non technical sources. Accordingly some of the methodologies used are hybridised, utilizing techniques drawn from both science and art. The three propositions examined in this thesis have been set out and a method to distinguish between the range of knowledge types found in the sources perspectives and features derived from the literature has been proposed.

2 Sources, perspectives and features for the PAT framework

“It might be argued that when playing the ‘imitation game’ the best strategy for the machine may possibly be something other than the imitation of the behavior of a man.” (Turing 1950)

In this chapter, the sources and perspectives for the PAT framework are outlined. The challenge of user evaluation of synthetic speech is positioned in the context of a number of other fields that attempt to understand or represent varied aspects of human speech. Each source or perspective can be categorised as one of three classes: additive, distractive or subtractive. Some perspectives resist a reductionist approach, offering something akin to a framework for this framework, while others posit usable metrics. The perspectives include robotics, linguistics, acting and performance, music, speech in the cinema, and ventriloquism. Silence and negative spaces are significant contributors to a number of perspectives and an analysis of these phenomena will contribute to several sections in this chapter, with a discussion at the end. Finally, based on the sources, perspectives and features discussed, a theoretical basis for the PAT framework is presented.

2.1 Breadth and depth

In the context of a computer science thesis, the breadth and choices of perspectives reviewed herein may appear unusual to the reader. Chapter 1 developed the notion that the design of synthetic speech artefacts is the design of a holistic entity that elicits a holistic response from the user. This is clearly demonstrated in numerous experiments by Nass & Brave (Nass & Brave op. cit). Accordingly, a broad-based approach to deriving appropriate perspectives from which a design framework may emerge seems to be required. As no other comparable multi-dimensional, interdisciplinary framework exists there is no extant agreement as to which perspectives should be included as theoretical support. This could lead to an issue of ‘breadth over depth’ if too many sources are thrown into the mix. To address this problem, this research

will focus on the two main disciplines implied by the thesis title: synthetic speech and all associate technological and scientific derivatives, including linguistics; and the performing arts, including acting and musical performance. This is aligned to the fundamental motivation for the research to determine whether the techniques derived from acting and performing are relevant when the actor or performer is not human. In addition to this, while a wide range of perspectives are discussed in this chapter, not all of them are strictly implemented in the system and studies. They are included for completeness and also to indicate potential for additional lines of research that could be exploited in future interdisciplinary research projects in synthetic speech design. Some perspectives requiring more detailed scrutiny, and all those subsequently implemented in the software system and the studies, are outlined in brief in this chapter and then discussed in detail at relevant points in the rest of the thesis.

2.2 Additive, distractive or subtractive processes

At the root of the problem of synthetic speech lies the difficulty in representing human-like features to other humans. People tend to be acutely aware of imperfections in attempts to represent human beings realistically. This capacity to detect imperfections is not the preserve of a subset of uniquely sensitive listeners; it applies to us all. Anecdotally, it may be fair to say that today, for most users, distinguishing between a recording of a real human voice and an artificial one takes a matter of seconds. The current solution to this problem is to attempt to make the voice sound more human by giving it more detail; this is an *additive* process. However, the problem of representing the human-like has not always been addressed by increasing the level of verisimilitude. In ventriloquism, for example, the ventriloquist's dummy is deliberately given physical features that distinguish it from a real person: a crudely articulated jaw, staring eyes, and small stature. The realism of the dummy is constrained, and the observer is asked to believe in something that is intentionally unrealistic. The ventriloquist chooses to concentrate on the features that are the most effective, and diverts the audience from those that are the least effective. Realism is not the point; hence ventriloquists have not, with the exception of Strassman (2009), adopted animatronics or other technologies to make the dummy more realistic. Similarly, in Shakespearean theatre, while the audience of the 16th century was perfectly aware that 'real' people did not speak in verse, the dramatist's art persuaded them to focus on the virtuosity of the actors rather than the believability of the

characters. Shakespeare's theatre depended on the audience's willingness to overlook unrealistic details, like the minimal sets and even the genders of some of the actors; it was accepted that the significant aspects of the drama could be rendered as effectively – if not more so - with the available technology ('technology' used in its broadest sense here to refer to the actors' bodies and voices).

Attempts to increase verisimilitude by adding more detail can be characterised as an *additive* processes. These processes can be contrasted with others, like ventriloquism or Elizabethan theatre, which we may call *distractive*, in that these processes seek to *distract* the audience from their lack of realistic detail. There is a third class of processes that we may call *subtractive*. In this class, the objective is to remove information to advantage the simulation. Somewhat contradictorily, this may involve adding new information that represents the new state. For example, the voice of R2D2 in Star Wars (Lucas, Kurtz, Williams et al. 1977) (CD Track 14) is rendered using expressive tones to contrast with the voice of C3PO, which is fully humanoid. R2D2 has an expressive voice (added) despite the absence of natural language (subtracted). This solution is further developed in the computer game 'The Sims' (Maxis 2009) (CD Track 15). In 'The Sims', the characters speak 'Simlish': a version of gobbledygook recorded by specially trained actors. The actors are able to emulate the emotional highs and lows of human speech without speaking a translatable language. In this case, the decision not to use natural language may be born of necessity - most likely cost - but the point remains that for these solutions to the problems of synthetic speech, a *subtractive* process is central to the overall aesthetic.

In the history of Western Art, the trend appears to be that initial attempts towards additive realism are followed by subsequent reactions in the contrary direction. The technique of 'perspective' in graphical art, for instance, is an example of an attempt to more accurately represent real objects in an artistic work; 'perspective' involves adjusting the size of a depicted object in order to account for distance, so that distant objects appear smaller and closer ones larger. This mimics the way the human eye perceives distance, and so attempts to enhance the verisimilitude of an artistic work to which it is applied. The obsessively faithful reproduction of surface realities had a short lived vogue in the form of superrealism or hyperrealism in the 1960s. Abstract art is the culmination of the opposite trend, where all but the artist's personal understanding of the essential is eliminated, resulting in a concentrated artefact with little or

no obvious relation to the real object that may have inspired it. The artist looks for the essence of the subject, and selectively eliminates the non-essential. In graphic design, 'white space', or 'negative space', serves something of the same purpose; graphic designers use unmarked spaces to emphasise the marked space, to focus the viewer's attention on the fundamental essence of an image. Movements away from realism may be due to a technological breakthrough that makes the effort less worthwhile (photography, photocopiers or 3D computer graphics take some responsibility), or due to a shift in audience expectation. As the accurate reproduction of real life becomes less difficult, it warrants less expense; the audience comes to expect the artist to invest as much time in personalizing, transcending, or distilling the subject matter as in representing it.

For a graphic designer, using 'white space' to represent negative visual space is largely unproblematic; for an animator, representing stillness - the negative space of animation - is troublesome. The difficulty of distinguishing between the appearance of stillness and the appearance of death is an issue in the design of humanoid robots. (Qazi, Wang & Ihsan-Ul-Haq 2006) cite MacDorman's analysis of the problem using Terror Management Theory (TMT) which equates unfamiliarity feelings with the degree to which the object reminds the viewer of mortality. In less demanding environments, such as avatar design, the animator has to be careful to maintain liveliness in the avatar when at rest for it not to appear dead, disinterested or broken. Pauses, silences or hesitance provide audible negative space and, in speech synthesis, are less well understood or exploited than their inverse. In other domains they are better utilised. For an actor, the dramatic pause is an affective device. In musical composition, expression is in part defined by the duration and location of rests, silences and pauses to contrast with sound. It is interesting to consider that in some spontaneous speech (contrasted with read speech), more than half of the duration of the speech may be occupied by silence. These auditory features all occupy the third class of processes (subtractive) and provide the principle processes of interest in this research.

The following classification of processes for synthetic speech enhancement is proposed:

1. Additive processes – Increase the detail of the speech sound to more closely match human speech sounds (greater verisimilitude).
2. Distractive processes - Provide other sounds or distractions so that the user is not aware of deficiencies in the synthesised speech.

3. Subtractive processes – Decrease the level of detail (less verisimilitude) and accordingly modify expectations.

These categories are not mutually exclusive. Some distractive processes - for example, rendering paralinguistic sounds such as laughter - are also additive. The process is distractive because, in this example, the paralinguistic laugh (used to signal good humor) distracts the listener from the absence of good humor in the prosodic rendering.

In synthetic speech research, refining the class of additive processes continues to be the approach of choice. In this research we plan to investigate distractive and subtractive classes. This is a rich vein of research and this thesis focuses on those with the most potential of presenting discrete metrics. In the following sections we describe sources and perspectives derived from the following fields:

1. In artificial intelligence (AI) and robotics, while there may be a need to capture deeper levels of human cognition, the simulation of shallower levels of human behavior can contribute to both subtractive and distractive models of synthetic speech production.
2. Investigations in linguistics have shown some comparatively simple models of paralinguistic variation that offer potential as models for distractive and subtractive processes.
3. The performing arts has evolved a number of strategies for economical (subtractive) representations of various facets of voice communication and expression.
4. The cinema presents a largely unexplored range of methods of all three classes of processes.

As has already been stated, the number of sources and perspectives examined over the course of this research is large. They have all been documented in this section. At the end of this chapter the sources and perspectives chosen for implementation in the PAT software tool are set out. However the PAT framework as distinct from the PAT software tool is informed by the full ranges of sources and perspectives reviewed in this chapter. Those that were not developed or tested in the tool are still significant in this broader context. Sources and perspectives that have made only a minor contribution to the tool or the framework but may still have potential are included here in order that they can be referred to unambiguously in Chapter 9 Conclusion and future research.

2.3 Artificial intelligence and robotics

Although the Turing Test (Turing 1950) was never intended as a test of synthetic speech communication,¹⁵ its position as the pivotal paradigm for computer intelligence expressed through language gives it relevance to this research. In the famous test the test participant would type and receive a response (via teletype) from one of two sources: an unseen human operative, or a computer. The participant should engage in dialogue with the system, unaware which of the two sources are responding. One interpretation of Turing's thesis¹⁶ is that for a computer intelligence to pass the test, the participant must remain unable to distinguish it from the human operative for the extent of the dialogue. Turing believed that passing this test would be evidence of sufficient machine intelligence to qualify as an equivalent to human intelligence. The test has yet to be passed. Much effort has gone into solving the problem additively by the application of more processing power, a larger database of knowledge, or increased interconnectivity, but before more options for brute force became available it spawned a number of related applications that apply the distractive tactics recommended above.

For example: in ELIZA¹⁷ (Weizenbaum 1966) characterised the computer respondent as a 'Rogerian psychotherapist'. As Weizenbaum says:

"ELIZA performs best when its human correspondent is initially instructed to "talk" to it, via the typewriter of course, just as one would to a psychiatrist. This mode of conversation was chosen because the psychiatric interview is one of the few examples of categorised dyadic natural language communication in which one of the participating pair is free to assume the pose of knowing almost nothing of the real world."

By casting the system with a credible role - the psychotherapist who is licensed to know nothing - Weizenbaum prevents ELIZA revealing her non-human source by demonstrating

¹⁵ Interspeech 2009, Brighton UK, is due to host a speech version of the Turing Test.

¹⁶ The tests can be interpreted differently as the original paper specifies that the test rests on the capacity of either a computer or a 'man' to successfully imitate a 'woman.' This detail does little to improve upon the more mainstream interpretation provided above.

¹⁷ Presumably the name is a pun on Bernard Shaw's character Eliza Doolittle in Pygmalion, who only repeats what her master says.

ignorance of things that should be known to a real human. Similarly with PARRY (Colby 1972), the respondent is cast as paranoid, thus licensing bizarre and irrational responses and creating a similarly credible, if extreme character. Both these examples demonstrate the power of creating a character that inhibits the expectation of the user within a specific, constrained framework. Neither application was designed to pass the Turing Test although a more recent example that pretended to be a human pretending to be a robot almost did¹⁸.

In Chapter 1 the notion of the ‘Uncanny Valley’ was introduced. Modeled as a steep sided valley Mori (Mori op.cit.) suggests that contrary to expectations, an over-realistic entity will, at a certain point, drive the user down the steep-sided valley to come to rest with the credibility of the entity at a lower point than less realistic manifestations (as depicted in Figure 4). The implications of Mori’s model are radical, with resonances in many different research areas. For example, in artificial intelligence research, underlying philosophical differences are exemplified in a debate distinguishing between two conflicting aspirations: ‘weak AI’ is intelligence intended to simulate human intelligence but not to possess its own human-like ‘mind’, while ‘strong AI’¹⁹ aspires to equal or exceed human intelligence, and so strives to possess human traits such as consciousness. By implication the ‘Uncanny Valley’ suggests that an AI embodiment that wishes to appear familiar to a user would be well advised both to be ‘weak’ and to announce the artifice it is employing, rather than concealing it. Failing to do so may result in the user being alienated. It is easy to imagine a similar problem arising in robotics as synthetic speech becomes increasingly life-like.

¹⁸ Reported on the New Scientist website 07:00 22 January 2009 by Colin Barras. “The 2008 winner of annual Turing test contest the Loebner Prize also won using a brilliantly simple strategy that Turing didn’t foresee. Elbot convinced three of 12 humans it was just like them by acting like a human pretending to be a robot.

¹⁹ The terms ‘weak’ and ‘strong’ AI were first coined in Searle, J. R. (1979). *Minds, brains and programs. Behavioral and Brain Sciences* 3, 417 - 457.

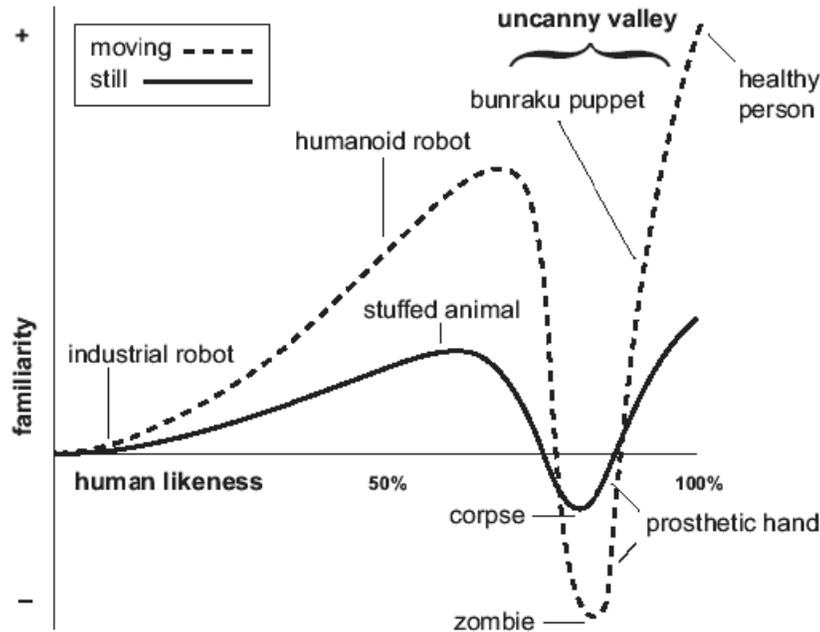


Figure 4: 'The Uncanny Valley.' (Mori 1970)

In Figure 4, 'familiarity' (the vertical axis) plunges as the artefact approaches 'human likeness' (horizontal axis). Mori also illustrates variation in the familiarity response between a static and moving embodiment.

The topography of 'The Uncanny Valley' is not universally agreed and some dispute its existence (Pollick 2004, Bartneck, Kanda, Ishiguro et al. 2007, Qazi, Wang & Ihsan-UI-Haq 2006). This research identifies topographic ambiguities, but it could be that the current direction of synthetic speech development – towards greater anthropomorphism - may still have to account for issues raised by Mori.

2.4 Linguistics - prosody

In terms of sources and perspectives for the design of synthetic voices, contributions from linguistics are of paramount importance. A detailed analysis of the specific linguistic concepts which inform the PAT software tool and framework is presented in Section 5.1. The following serves as a brief general introduction to the pertinent sub-disciplines in the field.

Linguists help define the underlying synthesis models by providing insights into the human systems that produce and modulate vibrations in the human vocal tract to produce intelligible speech. Linguistic research highlights the difficulties present in representing credible human speech in a machine by continually revealing more complex models of human speech production. Chapter 3 documents the usage of specific linguistic terms in this thesis in a structured vocabulary and states that the PAT framework is designed to modify speech at the most basic level (defined in the structured vocabulary as ‘neutral speech’) so that it is perceived by users as if it were speech at a higher level (defined in the model as a ‘voice’). Applying ‘prosody,’ defined by the title of Wennerstrom’s book as “The Music of Everyday Speech” (Wennerstrom 2001) to the flat intonation customarily found in speech synthesis might help users perceive it as a higher within the structured vocabulary. A speaker’s prosodic choices are one part of a complex interconnected environment of processes including lexical, contextual, emotional and unintended determinants, all of which act on the sound. Wennerstorm’s definition is enriched when we consider similar interconnectedness found in music between the rhythmic, harmonic and melodic determinants when combined with the performers’ ability, mood, intelligence, sensitivity and a multiplicity of less-obvious, but still potentially significant, variables. Many modern synthesis systems include a prosodic component; however, modeling prosody in speech synthesis particularly for a TTS system has proved to be a largely intractable problem. Rather like passing the Turing Test (described in 2.3), credible prosody seems to require a human-like intelligence and knowledge of context which is currently noncomputable²⁰. According to Mori’s principle, it seems possible that human-like prosody would be more likely to nudge the listener closer to the ‘Uncanny Valley’ than neutral or minimal prosody.

It may be possible to use other features derived from linguistics, commonly categorised as paralinguage²¹, as part of a distractive process, to make the user focus on a manner of speaking which can be delivered by the available technology.

²⁰ Noncomputable is defined either as time complex (i.e. taking a long time to complete) or space complex (i.e. based on the space requirements for data storage) Brookshear, J. G. (2000). *Computer science: an overview*, Reading, Mass.: Addison-Wesley.

²¹ There is no consensus in literature agreeing to categorise pauses or tempo variations as either prosodic or paralinguistic features of speech. Accordingly a composite form, paralinguistic/prosodic will be used in this thesis to describe modifiers applied to the synthetic speech stream in the studies.

2.4.1 Paralanguage

Although the major part of the prosodic system is governed by linguistic rules that ensure intelligibility, at runtime, the speaker, according to a very complex performance requirement, adjusts the rules. The adjustments are played against an underlying metrical phonology or hierarchical rhythmic structure that may be shared in a conversation situation and facilitates processing. It is paralanguage “...the variation of pitch, volume, tempo and voice quality that a speaker makes for pragmatic, emotional and stylistic reasons ...” (Wennerstrom 2001, p.60) that could provide a set of distractive features.

Phenomenon	Symbol
Pitch extremely high	+screech+ or even ++screech++
Pitch extremely low	fee fi fo fum
Volume	((louder / shouting / crescendo))
Quiet speech	° don't let anyone hear us!
Voice quality	((shrieks / mimics Groucho Marx voice / clenches jaw))
Sound effects	((whistles / makes trucks noise / imitates dog barking))
Laughter	hh: ha ha: huh: ah hah hah (transcribed to approximate sound)
Sounds from elsewhere	((siren goes by / slapping sound))
Rhythmic beats	/ beat / beat
Tempo goes up	>>
Tempo slows down	<<
Elongated syllable	...
Pause in seconds	(x.x)
Unmeasured micropause	(.)

Table 4: Symbols for paralanguage

Table 4 (ibid. p.61) makes it clear that the range of paralinguistic variables is large and that a synthetic rendering of some, more complex, paralinguistic variables may also have an impact on lower-level semantic functions.

In human speech, paralinguistic variation can be measured independently of the grammatical, lexical, phonological and intonational structure, but in a synthetic rendering some features could be confused with prosodic features whose usage could indicate a change of meaning rather than a ‘performance’ choice designed to improve the voice: to help users perceive it at a higher level in the structured vocabulary (see Chapter 3). This can be demonstrated by testing

the toolkits provided by some of the more recent synthetic voices (Loquendo 2008) (CD Track 5). Changes to the voice styles - e.g. assertive, friendly or professional - appears to have the effect of changing the meaning of the text. This is because the modified features, such as elongated syllables, volume or changes of pitch, are also part of the complex phenomena of stress used in speech to create or modify meaning (Crystal 1969, pp.113-120). While it is not possible to prevent any one paralinguistic/prosodic variable from acting as an unintended modifier of meaning, some features (pauses, rhythmic beats and sounds from elsewhere) are less intimately related to the semantic-prosodic system, and so may offer more flexibility. The scope to explore more of the available set of paralinguistic variables set out by Wennerstrom is discussed in 9.6 Further research.

As stated in the abstract a potential synonym for liveness is 'spontaneity'. The final linguistic sub-domain covered in this introduction, is research in 'spontaneous' speech: as such it provides the most significant contribution from the linguistic domain to the research presented herein. It is introduced briefly in the next section.

2.4.2 Spontaneous speech

Studies by F Goldman Eisler (Goldman Eisler 1968) into the characteristics of spontaneous speech identify a number of features. These may be briefly summarised as follows.

1. Spontaneous speech contains significantly more pauses than pre-prepared speech
2. The pauses fall in both grammatical and non-grammatical places
3. The pauses may be filled (by breath) or unfilled
4. The rate of speech is largely a function of the number of pauses and not the articulation rate
5. The speech rate follows a periodic pattern of slow-fast chunks
6. Significant individual variation between speakers exist

These features provide a particularly valuable model including metrics suitable as settings for the PAT software tool. Accordingly they are discussed in detail in section 5.1

The performing arts present a performer's mind, body and most significantly voice in the context of a display to others. Accordingly theorists and practitioners in the performing arts, in

particular musical performance and singing have developed some authority in discussions related to voice production and modification. This is the subject of the next section.

2.5 Music, the performing arts and cinema

The primary purpose of this section is to draw out those theories and practices which illuminate the processes that take place when a listener and a synthetic voice interact. The secondary purpose is to find those which directly inform the appropriate rendering of the audio produced by the speech synthesis system employed in this research. Frequently the two viewpoints - high level/theoretical and low-level/practical - overlap. For example, it is not possible to take the Renaissance playwright's detailed metrical system of pauses and simply apply them to a TTS reading. Some attempt to also render the context of Renaissance theatre may need to be made, including the language, the expectation of the audience even the rhetorical style of the voice. Teasing out the sources, perspectives and metrics from the performing arts that can be transported to the domain of synthetic speech and have a fair chance to be rendered with digital technology from those that can only provide a context for digital appropriation is challenging. The reader is advised to refer to Table 5 towards the end of this chapter for clarification.

2.5.1 What is acting?

*"The secret of acting is sincerity. If you can fake that, you've got it made."-
(George Burns)²²*

*"When someone laughs, then other men laugh back,
And when he weeps, then they weep in return. And so,
If you would have me weep, you must first suffer pain.
(Benedetti 2007, p.17. Citing Horace 'The Art of Poetry')*

It may be trivial to posit that the purpose of a synthetic voice is to 'act the role' of a human voice. It is unlikely that many in the field would disagree, however, that from this perspective a

²² This quote is variously attributed to Jean Giraudoux, George Burns and Groucho Marx

host of techniques that relate to the specific craft of acting are revealed and may be exploited. These techniques are obscured if we omit the word acting or substitute it for another that may imply a more profound similarity between the human voice and its synthesis in a machine. In acting, there is an actor and a character that the actor enacts. Despite WSOD, the audience does not lose the knowledge that both entities exist in simultaneous cooperation. As previously suggested, 'weak AI' (we could call this *acted* intelligence) and 'strong AI' (we could call this *being* intelligent) present quite distinct goals. The same may be true if we apply the weak/strong distinction to synthetic speech. The problem is that the terms 'acting' and 'being' have become somewhat confused in the field of acting theory, and this has repercussions in all fields that may wish to make use of the terms in order to draw distinctions.

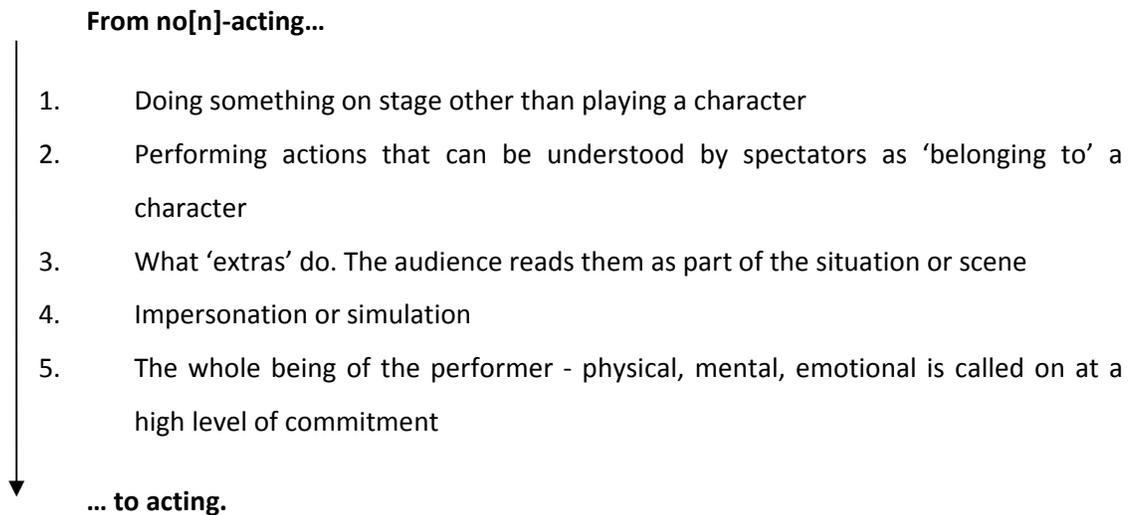
The confusion is the result of a widespread suspicion of the notion of pretence among acting theorists. 'Pretending' to be somebody else implies (to some critics) an approach to acting that trivialises the process, reducing it to a simulation of surface properties (defined as naturalism in 1.1.6) and thereby in some way missing the 'essential' properties of the character. This has led to a number of positions on acting that seek to strengthen the bond between the inner (psychological) life of the character and that of the actor (E.g. The poet Horace quoted at the beginning of this section). The actor is required to try to feel the same emotions as the character and then project these to the audience. As stated in the preface, various schools and techniques exist to help the actor with this process. Surprisingly this debate on the interpretation of the craft of acting is not new and has always been polarised.

Two thousand years ago, and representing the 'strong AI' side, Quintilian said "the prime essential for stirring the emotion of others is, in my opinion, first to feel those emotions oneself." (Roach 1993, p.24 citing Quintilian). In 1773, and representing the 'weak AI' side, Diderot says in 'Le paradoxe sur la comédien' "extreme sensibility makes middling actors ... in complete absence of sensibility is the possibility of the sublime actors" (ibid. p. 116, citing Diderot).

These apparent contradictions in the interpretation of the process of acting may lead the reader to question the values of acting as a framework in which to understand the role of the synthetic voice, but with closer examination the conceptual realignment brought about by accepting the paradox may prove valuable.

2.5.2 Finding rules for synthetic acting

A non-specialist, Western perspective on acting would probably coalesce around the ‘realist’ style of acting prevalent in most modern cinema in which the actor appears to be experiencing ‘real’ events and responding with ‘real’ emotions. But this is only one way to act. (Schechner 2006, citing Kirby), presents levels of acting from not-acting²³ to acting:



A suggested mapping for a synthetic speech actor on this continuum might be somewhere between point 2 and 4, thus avoiding the demanding requirements presented at the *acting* end of the scale. In order to avoid the Uncanny Valley, the synthetic actor will need to take care not to overreach itself and be sure to constrain the expectations of the audience. It would be convenient if a manual of acting existed that clearly defined the procedures required of an actor to present Kirby’s levels; a rule-set of acting. One of the objectives of this research is to attempt to uncover a set that could be applied to a synthetic voice.

Prior to the twentieth century, it is easy to find examples of attempts to codify acting as a series of rules for the actor to follow. However, as previously stated, the role of human intuition and nature in acting has always been hotly debated and there is little evidence of an imminent resolution. Roach provides a comprehensive overview of the debate (Roach op. cit.). In general, acting with the body, gestures and deportment has been subject to more rigorous

²³ Schechner misquotes Kirby. The wording used by Kirby is *non-acting*.

codification (Siddons 1822, Engel 1785), while the voice has had less attention except in the field of elocution²⁴ (Walker 1810, Burgh 1761).

There are exceptions in the twentieth century. Stanislavskii, paradoxically the primary architect of the psychological approach, was fascinated by opera and believed that the rules of the musical score should also apply to spoken theatre. Roach cites Stanislavskii in 'Building a Character' (Stanislavskii op. cit.):

“Rakhmanov, Tortsov’s assistant, introduces an “electrical conductor for plays.” This device consists of two silently blinking lights that could be placed in the prompter’s box and altered to blink like a silent metronome at different tempi as noted in the prompt box. Torsov and Rakhmanov demonstrate this device by playing scenes in accordance with tempi set randomly by the electrician, yet justifying each tempo with proper motivations. A gifted actor with proper training Stanislavskii believed should be able to respond to external tempo-rhythm without violating the law of inner justification” (Roach p.212 op. cit.)

More recently Samuel Beckett added what appear to be an annotated set of metrics prescribing the pauses in 'Not I' (Beckett 1973).

²⁴ Elocution is primarily concerned with ‘correctness,’ and, accordingly, tends to address speech production at a lower-level than that intended in this thesis. However, as a neglected and unfashionable discipline, it is possible that there may be some lost insights that could be of value to the field (see ‘9.6: Further Research’).

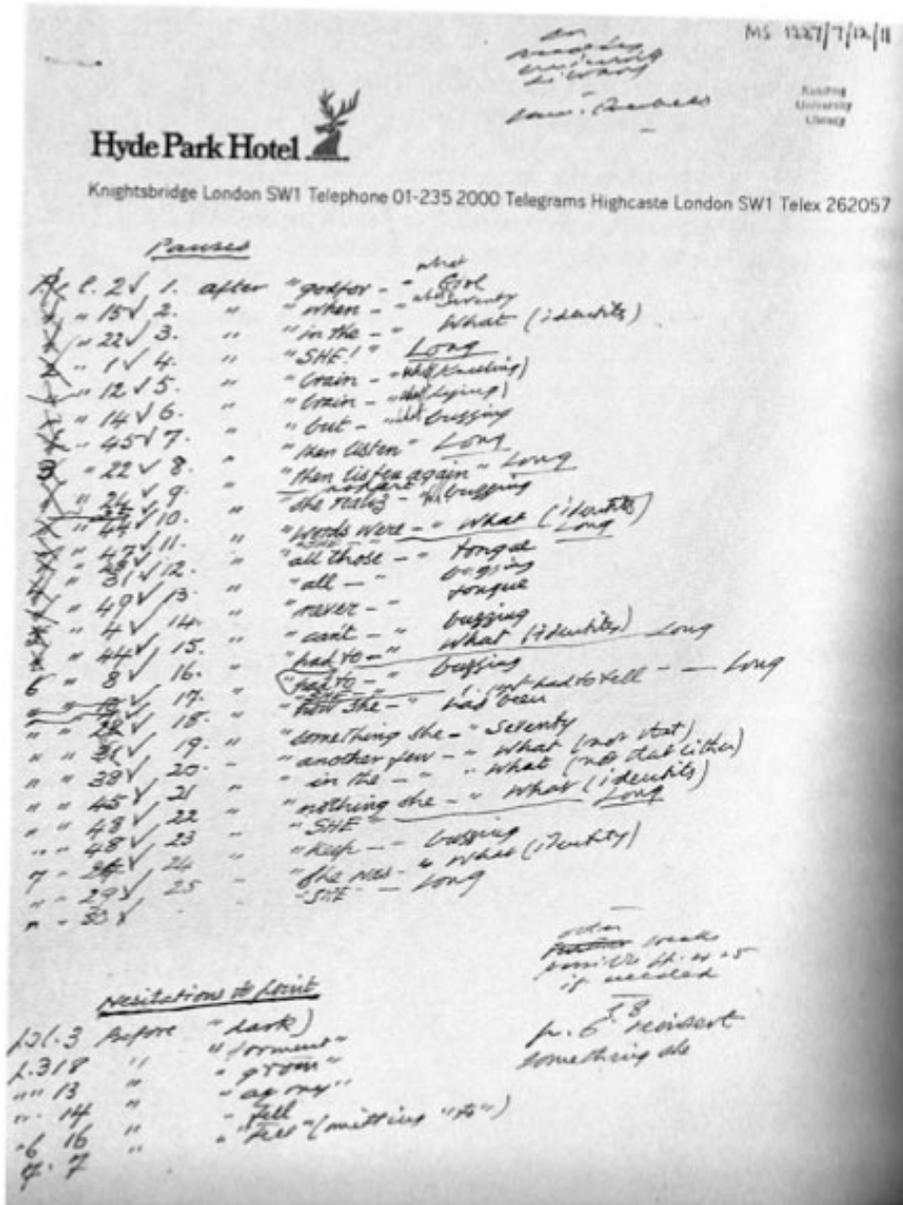


Figure 5: Samuel Beckett's notes for pauses in 'Not I' (Gontarski 1992)

Beckett's style owes much to the inherent musicality of language and, like Stanislavskii's 'electrical conductor', is designed to work in a context informed by musical traditions. As Wennerstrom has pointed out, the relationship between the underlying rules of prosody and those of music is intimate; thus the rule-set - the uncovering of which, as has been stated, is one of the objectives of this research - may be found in just such an intimate relationship between music and acting.

Despite a general dearth of useably precise metrics, the theatre has been successful in establishing an aesthetic ethos where the spectator is prepared to accept a less-than-complete, or imperfect, representation of reality. In so doing, it may be argued that the random metrics applied in Stanislavskii's electrical conductor could be useful. Although it will not be possible to justify each tempo with proper human motivations, it may still be possible to justify them using other means, particularly if the less than complete rendering encourages the audience to exercise WSOD. Indeed it may be argued that 'less-than-complete' is a crucial part of the experience of the performing arts as a celebration of the art of make-believe, in which some components of the drama are left to the imagination of the audience, rather than literally presented.

Some 20th century performance traditions have retained or rediscovered the 'less-than-complete' (subtractive) style of performance in which actor and spectator are both aware of the conscious application of craft or acting techniques designed to stimulate the imagination while not pretending to be 'real'. For our alternative synthetic voices to have a chance of greater appeal than the current solutions, achieving this separation of actor from reality is critical.

In 1908 Edward Gordon Craig wished to eradicate theatre of all the idiosyncrasies of human emotion and expression, and to create an Übermarionette (super-puppet) (Craig 1908). This mechanised vision was intended to provide theatre with actors whose refinement was determined by rules and spontaneity reduced to reflex conditioning. He quotes Flaubert: "Arts should be raised above personal affection and nervous susceptibility. It is time to give it the perfection of the physical sciences by means of a pitiless method" (Roach 1993, p.160 op.cit). Craig's position can be seen as a revolutionary reaction against the predominant trend toward psychological intuition, preferring a mechanistic approach. Perhaps the most well known contributor against the trend is Bertold Brecht (Brecht & Willett 1974) who set out to establish an acting method (known in English as 'alienation') that required the actor to objectify rather than embody the behaviour of a character. This allowed the spoken text to provide a commentary on the action of the character that could be examined by the audience. The character remained transparent to the audience with the actor and the ideas the dramatist wished to present always apparent. In the following extract, he captures something of the experience:

“The artist’s object is to appear strange and even surprising to the audience. He achieves this by looking strangely at himself and his work. As a result everything put forward by him has a touch of the amazing. Everyday things are thereby raised above the level of the obvious and automatic.”

(ibid. pp. 91-94)

Brecht appears to recommend a style of acting that rejects the everyday. Similarly, a synthetic voice should adopt a far-from-everyday style of acting and not feel obliged to represent the ordinary – which it is poorly qualified to do, anyway, being so far from ordinary.

Arguably, Craig and Brecht both hankered for the rigour and control that had been required for performances in Shakespeare’s time when, because of the design of the performance space the physical circumstances and human resources involved in the production, WSOD was mainly negotiated verbally and the actor’s ‘performance’ of the character, rather than the character itself, was the focus of the experience. Surprisingly these techniques derived from a theatrical tradition of 400 years ago, may have more to offer the contemporary synthetic voice actor than more modern techniques. An analysis of some acting techniques in the English Renaissance theatre is the subject of the next section.

2.5.3 Renaissance verse speaking.

When Shakespeare’s ‘two gentlemen’ enter at the beginning of many of his plays, they not only set the scene and introduce the main characters; they also establish a framework within which the audience’s imagination is to be exercised. They might as well say ‘This is what we provide for you: a prince, a young son, a good friendship, loyalty, but you will need to imagine the palace garden, the distant shoreline, the cave containing a library, the bear²⁵.’ This act of imagination was easy for Shakespeare’s audience because the convention was well established. The illusion only needed a hint of realism for the audience to flesh it out. Transparency (exposing the acting and performance processes) brought about by necessity is the hallmark of this type of theatre. For Shakespeare’s audience, acting was understood to be a transparent (or authentic) simulation of the actions of a person other than the actor. Although

²⁵ Actually, the bear may not have required imagination. There was a bear-pit next door to Shakespeare’s Globe theatre; it is thought that the famous stage direction in ‘A Winters Tale’, Act 3 Scene 3, ”Exit, pursued by a bear”, was facilitated by this geographical convenience.

they were expected to share the WSOD, they may also have enjoyed seeing through the cracks to the performer's technique. It may be posited that Renaissance acting operated in a tradition in which realistic representations were not possible, authenticity was a given, and prosody (particularly pauses) was an important modulator of 'liveness'.

Renaissance acting provides a surprisingly rich source for perspectives and metrics that assist in the development of a theoretical basis for the PAT framework. The rhetorical style of language was accentuated by contrasting the rhythmic rigor of verse with the relatively relaxed conversational style of the prose. Verse rhythm in English is built on the regular occurrence of prominent syllables. When rhythmic beats occur in speech, Wennerstrom categorises it as a paralinguistic phenomenon, and thus it may offer some potential for exploitation in the proposed set of paralinguistic/prosodic variables implemented in the PAT software tool.

Some scholars believe that there is evidence that the Renaissance playwrights encoded instructions for paralinguistic and prosodic phenomena into the script (Barton 1984, Hall 2004, Tucker 2002, Wright 1988)²⁶. The instructions were designed in part to assist the actor in keeping the audience's attention by ensuring that the important parts of the text were emphasised while less important parts could be 'glossed' over "trippingly on the tongue" - Hamlet Act 3, Sc. 2 (Shakespeare, Wells & Taylor 1994). Actors could interpret and execute these instructions on the fly with limited rehearsal and incomplete knowledge of the drama. In fact the system was so efficient that actors only needed to be provided with their own part and their cue in order to execute their performance. Who was who in the scene was presented on a written "platt" that could be consulted before they entered onto the stage. Thus much of the 'essential' information provided to a modern actor was poorly provided for a Renaissance actor but much more of the voice acting was readily accessible in the script itself.

Shakespeare's verse, for example, is written in iambic pentameters: ten syllables (or five 'iambs': alternating weak-strong units) per line of verse. 'If music be the food of love, play on;' (Twelfth Night. Act 1, Sc. 1) (ibid.) is an example of a perfect line of iambic pentameter.

If music be the food of love, play on;
 · _ · _ · _ · _ · _

Figure 6: A line of iambic pentameter. Weak beats are indicated by '·' and strong beats by ' _'

²⁶ Not all scholars agree with this analysis of Shakespeare's verse and the consequent implications for voice acting techniques. Space does not permit a more balanced review.

Iambic pentameter provided the basic metrical unit from which much of Shakespeare's verse is constructed. The breaths that the actors take tend to fall in rhythmically predictable places that may not always coincide with breaths taken for linguistic purposes (for example, to bring out meaning). The positions of pauses, and the types of pause (filled by a breath or not), are written into Shakespeare's code. The actor is usually free to choose the duration of the pause, although it is sometimes hinted at. Shakespeare's line length imposes a rule of 10 syllables followed by a breath and suggests the appropriate breathing space for the synthetic voice actor to capture the rhetorical style required by the language. Interestingly this 'rule' is confirmed by Goldman Eisler's experiments in speech/breath rates when comparing read and spontaneous speech:

"To maintain the same speech rate at a breath rate of 20 respirations/min, an output of only 10 syllables/breath is required which is not only easy to accomplish, requiring a lesser degree of voluntary control, but leaves the speaker with a surplus of expiratory air current to play with and use, so as to give his speech expression and rhetorical colour."

(Goldman Eisler op. cit. p.105)

Punctuation in Renaissance verse, although added later by editors and designed for reading, can be used to indicate a change in tone and provide an actor with a road map of tonal changes that give variety and (at least a simulation of) emotional range to the speech. Tone is best represented as timbre in synthetic speech, or the perceptual qualities which distinguish the same phoneme uttered with different spectral properties. A spectrogram will reveal the formants and harmonic frequency variations which define the individual quality of a phoneme. A change of tone in Renaissance verse was probably a hit-and-miss affair. Some of the vocal timbres (particularly those of the young boys impersonating women) would have been quite unnatural to modern ears. Actors were expected to match each other's tone even if an iambic pentameter was split over more than one line in order to maintain rhythmic coherence and to regulate the flow of the dialogue. In this respect verse-speaking mirrored natural discourse, in which speakers intuitively pick up one another's tone of voice.

Renaissance verse-speaking formalises a number of other prosodic/paralinguistic parameters, but these factors alone lead us to a tantalising possibility: can we improve a synthetic voice simply by applying the following three rules?

- After 10 syllables, breathe.
- If the breath falls in the middle of a word then wait until the end of the word.
- On a full stop pause and change tone.

An example of a human performance which exploits the full potential of the inherent instructions in the verse can be found on (CD Track 16). The same speech can be heard in alternative performances in (CD Track 17, Track 18, Track 19 and Track 20). In the latter cases, the actor and director have chosen to take more and more liberties with the original text in an attempt to imply (falsely, in the author's view) more emotion and realism. The techniques include changing the words, removing words, making vocal noises, adding diegetic and non-diegetic sounds (see section 2.5.10) and adding new voices.

2.5.4 Ventriloquism

The ventriloquial ('thrown') voice offers another framework for analysis of the human voice actor. Although today we associate ventriloquism with entertainment of a fairly tawdry variety, historically the voice of the 'vent' represented a kind of artificial intelligence. In the classical world, the Oracle of Delphi may have been presided over by ventriloquists, priests hiding in statues, and the user could interact with these 'artificial intelligences' by asking questions. The answers would be 'vented' back. We must presume that the hidden priests passed classical civilisation's 'Turing Test' - although the fact that they were pretending to be gods, rather than people gave them a similar advantage to 'ELIZA' and 'PARRY' (discussed in 2.3). In more recent history, users could communicate with other forms of intelligence, as spiritualists and mediums became adept at throwing the voice in such a convincing manner that clients would feel they were in the presence of beloved dead relatives. In both cases the objective was to create the illusion of not a 'real' being but a special being with particular powers.

Ventriloquism (pre-dummy) demonstrates the particular power of the disembodied voice. It is the power to be heard but not seen, that may be used to enforce the authority of gods. Connor, citing John Hull, says 'When we say that the divine being is invisible, we mean that we

do not have power over it. To say that the divine power was inaudible, however, would be to claim that it had no power over us.' (Connor op. cit. p. 24 citing Hull J.) Since the gramophone and the telephone (the arrival of which curiously coincided with a revival of interest in spiritualism and mediums), we have become used to disembodied voices; however, the human impulse to associate a voice with a legitimate source has not diminished. By legitimate source we mean a source that could give rise to the voice in accord with Connor's 'cultural sensorium' or Rick Altman's 'sound hermeneutic' (Altman 1980). As Connor says:

"To understand the operations of ventriloquism, in the larger sense of the separation of voice (and sounds) from their source, and the compensatory ascription of source to those sounds, is to go a long way towards understanding the construction and transformation of what may be called the cultural sensorium, or the system of relations, interimplications and exchanges between the senses" (Connor p.22 *ibid.*).

Readers may experience a feeling of illegitimacy, or conflicted 'cultural sensorium', when viewing a TV advertisement where the actor's visual appearance and voice do not appear to fit together despite attempts at accurate lip syncing. The legitimacy of the source lies at the root of the problem in speech synthesis. If a human sounding voice cannot be associated with a legitimate source then it will be perceived as uncanny.

The solution adopted by ventriloquism is the dummy. By associating the voice with a puppet the audience has the legitimate source they need and the ventriloquial act can be transformed from the uncanny into entertainment. This may indicate the possibility for a relatively reassuring effect that a physical embodiment, cartoon or avatar is likely to have on the experience of users experiencing synthetic voices. If the voice can be ascribed to a cartoon duck or robot then it may no longer be uncanny. Additional work beyond the scope of this thesis is required to verify this assertion beyond speculation. However, the point remains that in many application domains (e.g. telephonic, mobile and public address) a visual embodiment is not an option.

Although the distractive tactics employed in the dummy may not be available, the ventriloquial voice itself does provide some interesting perspectives. As a voice, it is adapted to the special circumstances of use and practical constraints. It is frequently cast as a low comedy character, where perfect diction is not expected. The peculiar noises resulting from the mechanics of

voice production employed by the ventriloquist, including lisps, slurs and squeaks, are acceptable in such a character, and consequently, as with 'PARRY' (see 2.3) the voice fits the context. If we accept the hypothesis that we may be able to empower a synthetic voice with the qualities of a special being (e.g. a machine imbued with liveness, rather than the qualities of a human being) then ventriloquism provides us with an interesting framework in which to imagine alternative synthetic voices suited to the context in which they are most often encountered. The notion of the ventriloquial voice is an important contributor to the culminating PAT framework test (see 7.1.5).

Historically, opera has its roots in the grandiose entertainments of the court. In such an environment the subject matter was expected to be lofty, rather than low comedy, very often involving the direct engagement, through the 'deus ex machina', of a supernatural force. This class of unreal voices is the subject of the next section.

2.5.5 Singing and opera

In considering a mechanised approach to speech production, opera is very relevant. Not only is the form highly theoretical and therefore supported by a significant body of research, case-studies, metrics and notational rigour, it is frequently required to traverse a challenging domain between the fanciful and the realistic. Opera composers have been asked to find ways of presenting shades of speech that occupy a wide range of positions on a continuum that extends from natural speech to sounds that are barely recognisable as human-derived. The solution has been combined in a range of musical and stylistic forms to produce the right degree of 'real' and of 'unreal'. In this respect, they operate in the same domain as that intended by the designer of synthetic speech.

In the evolution of human speech, some argue that *singing* came first (Barker 2004, p.68). This seems reasonable when we consider that the reflex sound made at a human birth is much closer to the sounds we call singing than those we call speaking. It is curious that a sound that appears to be entirely natural has come to be perceived by many as unsuitable for natural expression. Were we to disassociate singing from the discipline of music and use other, less certain, terms - chanting, keening, or even rapping - then this impression may begin to diminish. There is a wealth of different singing styles distributed across a similar wealth of

global musical styles, many of which would be unfamiliar to Western audiences. Unfortunately, investigating all of these and their relevance as potential paradigms for the design of synthetic voices lies outside the scope of this thesis²⁷.

In the film 'Mamma Mia' (Craymer, Goetzman, Johnson et al. 2008), the actors were required to adopt a style of expression in speech which some might regard pejoratively as 'over-the-top'. We may suspect, however, that the colourful prosody was a deliberate ploy to bridge the transition between normal speech and singing; a transition which might have otherwise sounded uncanny to the listener. Barker cites Stockhausen and provides a continuum between speech and song thus:

1. *Parlando*²⁸, where tone and speed imitate colloquial speech
 2. *Quasi parlando*, where the curve of the speech oscillates, but with fixed duration of syllables and intensity
 3. *Syllabic song*, where all the parameters are exactly proportionally set
 4. *Melismatic song*, where the musical parameters become dominant and tones predominate over syllables
 5. "Pure music" performed with the mouth closed
- (Barker 2004, p.68)

Opera composers developed a similar set of options when changing operatic conventions demanded characters that operated at a prosaic, day-to-day level, as well as the more fantastic level of song. The solutions reproduced from Moore (1962, p.583) came in a number of different forms:

- *Recitativo*: (CD Track 24) A passage of dialogue which is sung in a style more or less like prose speech, and which may be accompanied by anything from a simple continuo section – consisting perhaps of only a double bass and a harpsichord – playing a few spare chords to a complex web of fully orchestrated counterpoint.
- *Recitativo misurato*: (CD Track 22 and Track 21) A recitative supported by measured music instead of the isolated, rhythmless chords of dry recitative.

²⁷ The author is in receipt of a Wingate Scholarship to pursue this line of research in the future.

²⁸ "Parlando" means 'speaking' in Italian

- Recitative secco: (CD Track 23) One step removed from natural prose speech. The singer intones the words with harmonically determined notes, but not yet with anything that could be called melody.
- Sprechstimme: (CD Track 26) Translates as the ‘speaking voice’; hence, music which is written to be sung in the manner of someone speaking. ‘Sprechstimme’ designates, particularly, a technique of imitating tones and rhythms of speech with music, which permits the composer to direct the dialogue with absolute accuracy, as well as to accompany it.

Some 19th century Italian operas address the problem of dialogue by never allowing the singers to speak. Curiously, these works have been labeled ‘verismo’ (realistic), despite the usually-melodramatic subject matter. In ‘verismo’ opera, the characters are no longer obliged to express themselves in a poetic language or heightened style; rather they speak an “ordinary common, or idiomatic, language with no regard to metric regularity.” (Rizzo 2008). The orchestra supports the words with bold melodic sweeps “giving it an oral character, almost one of direct speech –gestural rather than abstract or logical.” (Mallach 2007). Various confluences of techniques appear in opera and musical theatre which may variously be described as melodrama or underscored speech (CD Track 25).

Speech-like forms in opera are an attempt to smooth over the join between speech and song thus producing a more natural effect. Extended vocal technique emerges from the same milieu but with the objective of driving the human voice toward sounds that are not found in standard speech but which may have additional expressive possibilities.

2.5.6 Extended vocal technique

Some 20th Century Western composers have incorporated unusual vocal sounds or ‘extended vocal technique’ (EVT) into vocal compositions (CD Track 31). While it may not be strictly accurate to describe EVT as providing a musical paralinguage, within the context of this research, EVT utterances add to the paralinguistic repertoire available to the synthetic voice. Paralinguistic utterances are part of a large repertoire of EVT sound-objects that extend the human voice to focus on the less-usual sounds it is capable of producing. Many of these sounds have been incorporated into linguistic or artistic expression in non-Western cultures, but in

Western music and linguistic expression, they remain a rarity. Wishart (Wishart 2002) records categories of 165 samples of individual sounds produced using the human voice. To give a flavour of the range of sounds and sources for sound, the categories he posits are reproduced below:

- The glottis and windpipe
- The tongue
- The lips and cheeks
- Filters
- Filtered noise and whistles
- Double and treble production
- Air stream and other effects
- Water effects
- Transformations (exhaled sustainable sounds)
- Inhaled sounds
- Pulses
- Voiced pulses
- Pulses with stops, buzzes, filtering
- Simultaneous and alternated pulses
- Pulses (clicks and combinations)
- Transitions and percussives
- Multiplexes and complex articulation

(Wishart 2002, pp.345-351)

Some of the items on Wishart's list correspond to the paralinguistic sounds outlined by Wennerstrom (see Table 4) as well as supporting the notion of a structured distribution of breath-like interruptions in spontaneous speech and renaissance verse, to enhance expressiveness. In musical vocal expression, breathing is an integral part of the structures created by composers, as well as the process of reproduction delivered by performers. Audible breathing in musical speech is generally discouraged, but in EVT several composers have exploited the expressive powers of different types of audible breathing (Anhalt 1984, pp.209-214) (CD Track 32). The refinement and granularity found in Wishart's recordings, and the

creativity displayed by composers and performers in integrating audible breathing into performance, supports the view that breathing may be exploited as an expressive enhancement to liveness in the synthetic voice. Some work has already been done on this (Whalen, Hoequist & Sheffert 1995), although the focus has been limited to improving comprehension rather than expressive potential (see Section 5.1.2).

Theoretically, expressive breaths and sounds from Wishart's list might be used to add liveness to a synthetic voice. Although the sounds have a natural human source, when they are isolated from the normal speech voice stream and emphasised through amplification or repetition the effect is likely to be expressive but unnatural, and thus may accord with the objective to render liveness without reference to realism. The rendering of these sounds may be an additive process but as Whalen et al suggest it may lead to an improvement in the perceived acceptability of the synthetic voice.

Acceptability and expressiveness are part of a complex vocabulary of descriptors for synthetic speech. These are dissected in the next chapter but for now a brief introduction to the computation of musical expressiveness is the subject of the next short section.

2.5.7 Musical expression

In the field of the communication of emotion in musical performance, Juslin (Juslin & Sloboda 2001) proposes a formal distribution of some of the mechanisms of music in expressing emotion. He explores the five most studied emotions (happiness, sadness, anger, fear, love/tenderness). Using data from other studies in which participants rated the valence and activity of emotional terms, he plots the placement of each emotional expression to a set of precise acoustic descriptions of the musical mechanisms. For example, sadness is mapped to low sound level, low activity and negative valence (see figure 7).

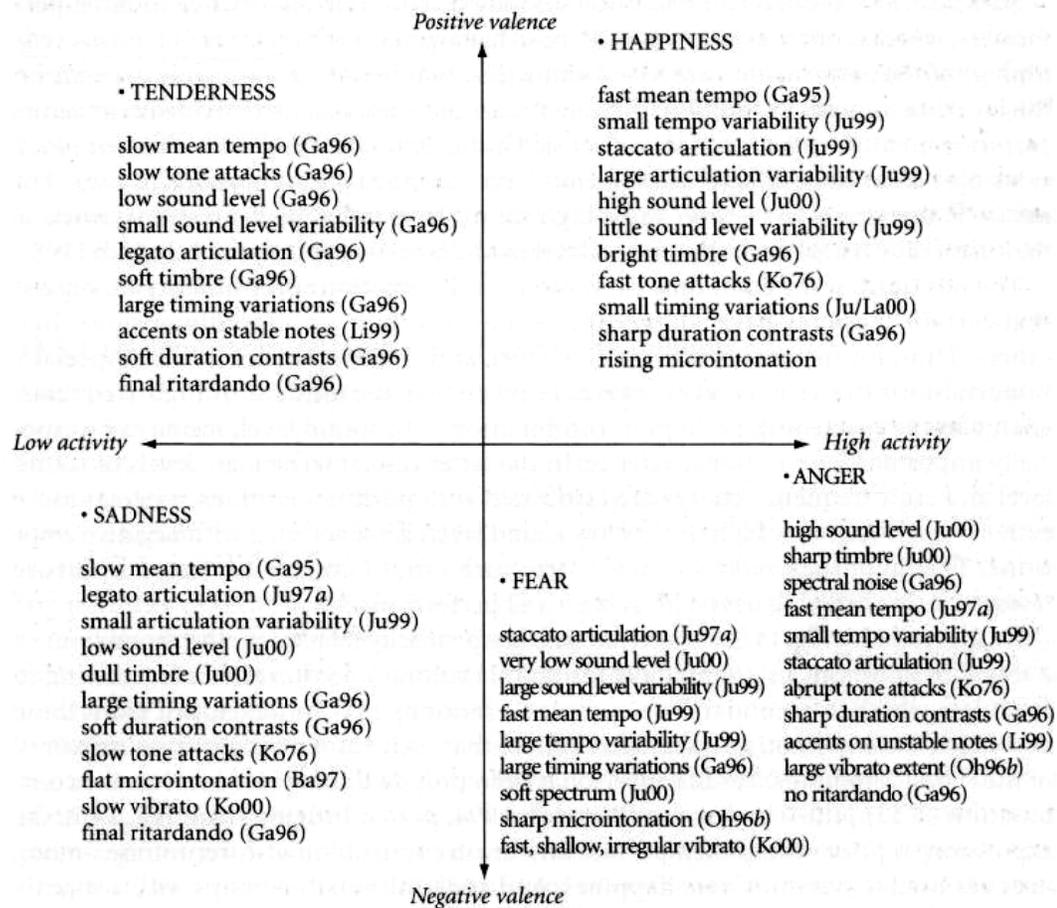


Figure 7: Plotting emotions to musical mechanisms from (Juslin & Sloboda p.315 *ibid.*)²⁹

Juslin states:

“The single cue combination that was most expressive according to listeners had the following characteristics (with cues in order of predictive strength): legato articulation, soft spectrum, slow tempo, high sound level and slow tone attacks. This raised an interesting question; is this how performers should play in order to be judged as expressive by listeners?”

(Juslin & Sloboda p.317 *ibid.*)

²⁹ The codes presented in brackets in this illustration refer to the cited studies in the original article. Readers should refer to the original article for clarification

This suggests that similar formulas may exist for a range of other effects; perhaps including complex effects such as liveness. Juslin's findings are supported by Widmer (Widmer 2002). He reports surprisingly good results from experiments in machine recognition of individual expressive playing styles by famous classical music interpreters. This is despite the fact the performance measures are extremely crude, comprising only beat-level tempo and beat-level overall loudness.

This evidence suggests that simple manipulations to some basic expressive parameters, amplitude and speed may serve to propagate the liveness envisioned in the PAT framework.

2.5.8 Humanisation algorithms and randomness

Electronic musicians using MIDI may choose to compose manually by entering values into a software editing environment rather than playing the notes using a keyboard. On playback, the result may sound mechanical, lacking the nuance that occurs naturally when playing an instrument. Algorithms have been developed to add humanness to the musical data and these take a number of forms, usually with a random component shown in List 1.

1. Perturbations to the temporal alignment of notes to the underlying pulse.
2. Variation to the attack or velocity applied to the notes.
3. Variation to the length of notes.
4. Variation in tuning and pitch.
5. Phrase randomisation.

List 1: Humanness algorithms

To illustrate humanisation (CD Track 27 to Track 30) demonstrates a simplified version of the process when applied to a crudely executed machine performance. This is also illustrated in Figure 8 to Figure 10.

On (Track 27), illustrated in Figure 8, a human performance of four bars from Bach Cello Suite No. 1 can be heard on a real cello. On (Track 28), illustrated in Figure 9, a typical machine performance of the same piece of music can be heard³⁰ on a synthetic cello. All the notes are

³⁰ The quality of the synthetic cello sound is not relevant and should be ignored in these examples.

played with exactly the same velocity and exactly in time. On (Track 29) each note is positioned to a random value between 0 milliseconds (ms) to 30 ms either side of the beat. On (Track 30), illustrated in Figure 10, additional humanisation is applied by randomly varying the velocity (loudness) of each note by between 50% and 100% of the standard velocity heard in (Track 28). The examples have been executed crudely for the sake of clarity but the same principles may be applied (with far more subtle results) using more advanced techniques and less-obvious changes.

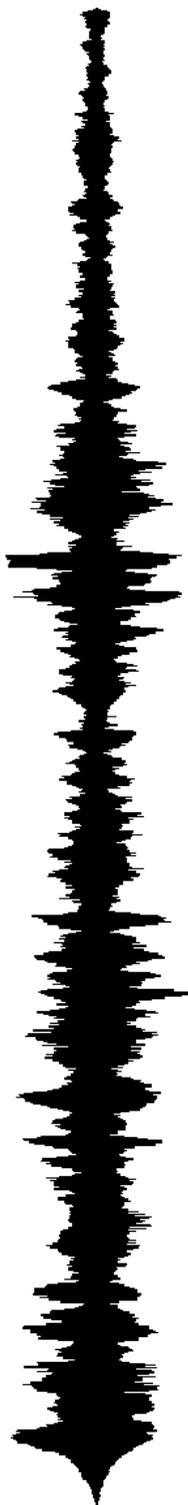


Figure 8: Waveform for the human performance

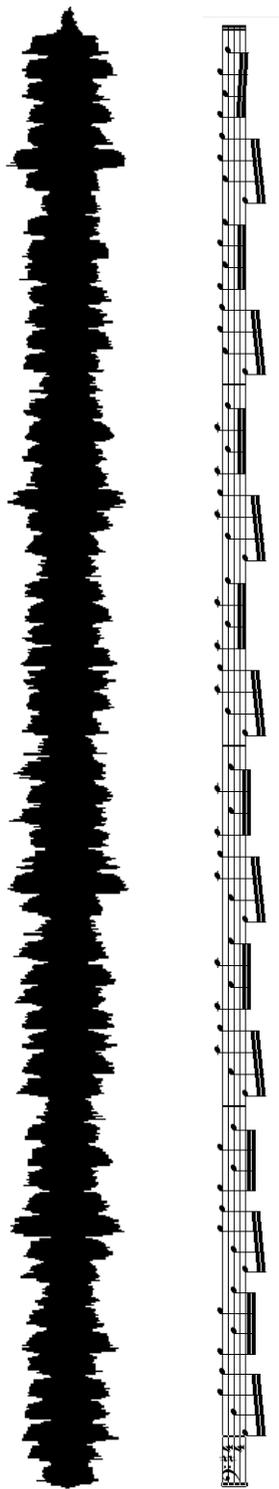


Figure 9: Waveform and notation for the machine performance

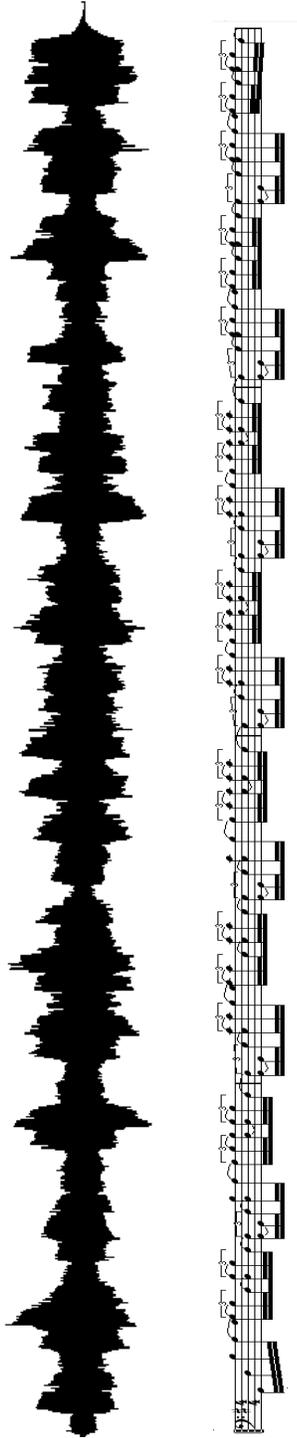


Figure 10: Waveform and notation for the “humanized” machine performance

Items 1, 2, and 3 in List 1 are implemented in a set of 'DNA Groove Templates' (Cholakis 1992) marketed by 'Numerical Sound'. The marketing materials say that:

"DNA grooves represent the human touch that makes music interesting and compelling. DNA grooves bring the feel of live music into the realm of MIDI sequencing."

Interestingly, Cholakis's method is to extract the information for the template from live performances by famous musicians. This contrasts with the templates provided in most well-known sequencing packages that are based on providing user control over random parameters. More elaborate humanisation algorithms are required in packages either processing or synthesising singing. Random perturbations provided in Voice Live (TC-Helicon 2008) is made up from a combination of Flextime™-based time randomisation, pitch randomisation, and pitch inflection, also known as 'scoop'.

The randomness programmed into these applications is constrained by heuristics, and there may be additional constraints applied to the randomness of the parameters generated by constraining algorithms such as filtered randomness (Rabin 2004). The software provides a means of controlling these parameters, usually with an expectation that the user will iteratively modify the settings until the optimum (most-human) setting is achieved.

This process is better expressed by the less-well-known term 'aleatoric'. Werner Meyer-Eppler describes a process as aleatoric "... if its course is determined in general but depends on chance in detail" (Meyer-Eppler 1957). The significance of randomness in creative systems is the subject of debate (Boden 1994, Boden 1995, Perkins 1994). Although there are plenty of anecdotal examples (the serendipitous discovery of penicillin, Bakelite and Velcro) that may have been accelerated by chance occurrences, Perkins argues that the 'ground has to be prepared' - a more structured process of research and experimentation must take place - for randomness to prove useful. Another way to describe this is that we have to be ready for randomness if we are going to make the best of it. Turing imagines random processes contributing to evolutionary and learning processes in the emerging digital technology of the time (Turing op. cit.). It is reasonable to suppose that the ability to assimilate randomness and transform it into serendipity (Campos & Figueredo 2002) would be desirable. Unrestrained

randomness could be described as the opposite of creativity, but restrained randomness has been used in the arts as a stimulus and synthesis of spontaneity. Examples include 'Action Art' - canvases daubed with random and spontaneous smears of paint - and aleatoric music such as John Cage's 'Music of Changes' (Cage 1961) and Mozart's 'Dice Music' (Mozart 1787).

Randomness has been used to stimulate unexpected solutions during the creative process (Eno & Schmidt 1975), as well as to modify systems to provide a human touch or an illusion of greater spontaneity. The paralinguistic/prosodic modifiers implemented in the PAT software tool (see section 6.2) are, to varying degrees, subject to randomness. In the studies, users preferred random settings than predetermined settings for some variables. Randomness provides a simple method of suggesting spontaneity, which may be applicable to the paralinguistic and prosodic modifications proposed in this research.

'Liveness' is the subject of this research, as established in the title. Of all the sources and perspectives discussed in this chapter, 'liveness' is the most abstract and ambiguous. With its roots in performance theory, it may appear some distance from the solid ground expected to support a computer science thesis. To help the reader, 'liveness' has been stripped down to only three concepts: place, authenticity and time. These concepts are the subject of the next section.

2.5.9 Liveness

Auslander's authoritative work on 'liveness' is in part intended to critique the special place occupied by 'live performance' when compared against mediatised (technically mediated) performance. He seeks to debunk descriptions such as 'the magic of live performance', and reminds us that the word 'live', in the context of performance, did not appear in dictionaries until the 1930s, when the converse, through recording, was well established. Thus the status of 'live', meaning 'real', and 'recorded', meaning 'virtual' is not clear cut. Through examples such as the ubiquitous use of amplification, video screens and audio sampling, he argues that the distinction between live and mediatised is now so imprecise as to be virtually meaningless. Thus it may be argued that a comparable distinction that could be made between a 'live' synthetic voice and a 'non-live' synthetic voice may need closer examination.

2.5.9.1 Authenticity

The value of 'liveness' is represented by Auslander's term 'authenticity'. 'Authenticity' may be understood as the subjective value the listener ascribes to the performance experience. The example he gives is going to the Shea Stadium concert by the Beatles, or viewing a live recording of the same concert. The former experience has greater authenticity, even if the definition and fidelity (in the conventional sense of high-fidelity audio reproduction) of the subsequent recording is better. However, authenticity is not an absolute associated with proximity to the human original. Grades of authenticity exist, and can be manipulated. A bootleg recording of the original concert made some distance from the stage and largely saturated with audience noise would also have greater authenticity than the live feed from the recording console despite the fact that both the physical proximity to the artists and the audibility of the artists is reduced. Auslander provides another example that may help clarify this ambiguity. In the 1970's analogue technology was developed that allowed pop musicians to sequence computer-generated musical sounds into songs. Despite the fact that, in performance, these sounds required no human performers, and thus had little authenticity, it is now the case that – with the popularisation of 'sampling' by musicians; a technique in which segments of other pieces of music are integrated into a new composition - modern performances and recordings that make use of this 'inauthentic' technology. Employing these 'old' sounds in 'new' compositions would now be considered as authentic. What an individual user considers authentic changes over time. This ambiguity has given rise to a situation by which the degree to which artists disambiguates the inauthenticity of their performance perversely gives rise to a higher authenticity-rating by their audience³¹. In evaluating authenticity it appears that context is critical. For a synthetic voice to be valued as 'authentic', the context in which it is experienced is similarly critical. In our extrapolation from Auslander we have henceforth called this factor 'place'.

2.5.9.2 Place

To optimise 'liveness' the performance place and the audience place have to be the same. This is simply illustrated by referring again to Auslander's example. For the audience within the Shea Stadium, the performance by the Beatles had undisputed liveness. A live radio broadcast

³¹ For example, when, on the BBC programme 'Top of the Pops', Liam and Noel Gallagher of 'Oasis' swapped roles, making apparent the fact that they were miming to a backing track, it signalled a trend of 'authentic' artists who could deliberately flout the convention of miming.

heard at home would have less liveness. Were this the only quantification of place available, then, as synthetic voices do not occupy physical space (having no physical embodiment), 'place' would have no relevance. However, synthetic voices occupy an auditory 'place' that can be either shared with the user's place, or be deliberately apart from it. If the notion is that a shared 'place' is also a shared context then the place can also exist cognitively but not be required to exist physically. This two-pronged approach to 'place' as both auditory and contextual reveals two rendering strategies. At the level of the first prong – the auditory level- 'place' can be rendered using a range of tools commonplace in radio and film production. These will be discussed in further detail in the section 2.5.10, but for now, at its simplest, the voice can be set within a soundscape that locates it in a particular place. For a TTS screenreader reading 'Wuthering Heights', a windy moor would be an appropriate soundscape. The physical place is unlikely to be shared with the user, unless they happen to be on a windy moor at the time of listening, but the cognitive context is clearly shared. The user is imagining the story as set on the moors, and the voice is situated in the contextually appropriate place, even though neither is physically there. The second prong - context - in this case depends on the script provided to the screen reader which may describe the intended place or evoke it using other literary techniques. Thus the script is a very significant player in determining the correct rendering of place, and hence, as with a stage play, the script in the PAT framework helps to construct the component 'liveness' even when it seems none can physically exist.

2.5.9.3 Time

In order to optimise 'liveness', the performance's location in time and the audience's location in time, have to be the same. 'Time' is the final dimension extrapolated from Auslander's theory. By now the reader may have recognised a pattern in the way 'liveness', which is essentially concerned with the difference between the real and the recorded, can be reframed to deal with the real recording and the unreal recording, thus providing an exact analogue with synthetic speech implementations. By locating the shared element of time in both the script and the audio rendering, the sense of a shared real-time experience may be projected. Another solution would be to devise a speech production system where real-time processing occurs and communicates the essential 'realtime' of the process to the user. This would probably mean allowing errors and other indicators that the speech events were not subject to advance processing to be passed through. But Auslander makes it clear that this is not necessary in order to optimise liveness and that the illusion of 'realtime' is all that is required. He cites the

case of 'Milli Vanilli' whose Grammy award was criticised when it was discovered that the duo had not only lip-synched in their concerts but had never sung on their albums. In this case, the real-time illusion had either fooled their audiences or (more likely) their audiences were quite aware of the deception but were happy to be fooled – suspending disbelief. In performance, the belief that time is passing at the same rate for the performer as the audience is often deliberately subverted by miraculously fast costume-changes, changes of set, or, particularly in film, by flashbacks. The critical thing is that in most cases the time frame, presented by the performers and accepted by the audience, is clearly understood. To be specific: a synthetic voice that says 'how are you today' is pretending to occupy the same timeframe as the user, and thus is required to step out of the role of performer and occupy the role of 'a being', comparable in all temporal respects to the user. With the technology we have today, the 'synthetic performer' is sure to come unstuck or to stumble down the 'Uncanny Valley'.

Although the rendering of place, authenticity and time are not the explicit objectives of the film soundtrack, which is primarily there to support the film visuals, it may be suggested that sounds, music and voices can be used by filmmakers to manipulate the audience's sense of liveness. This provides the subject of the next section.

2.5.10 Cinematic sounds

Sound in cinema has received significantly less attention by researchers than the visuals. This corresponds to the public perception of cinema as a visual art with an 'accompanying' soundtrack. Although primarily concerned with the audio-visual, Michel Chion (Chion & Gorbman 1999, Chion & Gorbman 1994) sets out a rich framework in which the interactive processes operating on the film viewer at the auditory level are explored. A number of these are processes that could also be said to apply in interaction with a synthetic voice and to our extrapolations from Auslander's theory of 'liveness'.

The first distinction to be made in considering forms of film sound is between *diegetic* and *non-diegetic* sounds. Diegetic sound can be said to occupy the same place as the action on screen. The sound of running water with an image of a tap being turned on would be an example. Non-diegetic sounds belong in some respects to the convention of the pit orchestra that accompanied silent movies. Music is usually non diegetic, although if it is heard while an image

of its source is shown on screen it is diegetic, conforming to Rick Altman's 'sound hermeneutic' (Altman op. cit.). Sound designers can make subtle use of these forms by moving the audience's perception along a continuum between these extremes. For example, in a romantic musical it is not uncommon for a scene to begin with an image of a wind-up gramophone in a scruffy kitchen, accompanied by scratchy sounds with low frequencies attenuated. Then as the scene continues and the romantic couple begin to dance the low frequencies are added, full 5.1 Dolby sound is relayed by the auditorium speakers and the absurdity of this sudden apparent upgrade to the on-screen audio system is quickly forgotten.

Voices in film are normally used diegetically, with careful attention placed on ensuring that images of lips synchronise with speech³². However, voice-overs are normally non-diegetic, occupying an acoustic space at-odds with the screen image. Examples of diegetic and non-diegetic sounds are to be found on CD Track 33. In this track, an excerpt from the famous last scene of *Psycho* (Hitchcock 1960), the ebb and flow of diegetic and non-diegetic sounds may be heard.

1. The excerpt begins with the diegetic sound of the police station door slamming and the footsteps of the officer.
2. The voice of Norman Bates' Mother saying "thank you" is diegetic; although not seen, it is heard through the open door of the cell.
3. The subsequent speech by the mother is accompanied by a close-up of Norman Bates implying the voice is in his head and therefore non-diegetic, this is reinforced by the non-diegetic string sounds in the musical score.
4. The last moments of the movie show an image of a car being dragged out of the swamp. The diegetic sound of the chain is heard accompanied by powerful non-diegetic music finishing on the sort of dissonant chord frequently associated with horror.

At this point, the reader may question the value of theories derived from a visual medium when no visualisation is envisaged in the PAT framework. However, if we refer back to the previous section and consider the two-pronged cognitive and auditory approach posited in the simulation of liveness, the relevance of film sound may be clearer. It is possible to create a non

³² Some European nations are so used to relaxed (imprecise) lip synchronization, the result of over-dubbing the dialogue, post filming, that they complain that it is unrealistic if 'live' recorded sound is used.

diegetic-sound object without a visual element if the visual element is rendered cognitively and not visually. An example is in one of the PAT framework tests. In the test, the synthetic voice lamenting the passing of time described the waves breaking on the sea shore. The non-diegetic sound that accompanied the description was of a mantle clock. The non-diegetic sound rendered a scene at odds with the literal location (clocks are not customarily heard at the sea-shore), but it was hypothesised that liveness would increase as the evocation of the temporal values reflected in the script and shared with the listener would be more effective (see section 7.1.2).

Space constraints do not permit a detailed account of all the insights presented in Chion's work; however a brief summary of the key points is of value.

- Rendered sound: sound that appears 'more real' than the real thing; for example, walking on cornstarch sounds, to the audience, more like walking in snow than the actual sound of someone walking in snow.
- Synchresis: a composite formed of the words synchronisation and synthesis, meaning "The spontaneous and irresistible mental fusion, completely free of any logic, that happens between a sound and a visual when these occur at exactly the same time." (Chion & Gorbman op. cit. From the 'Forward' by Walter Murch. p. xix)
- Empathetic and anempathetic effects: sounds that appear to reinforce the sentiment being expressed, or that deliberately demonstrate indifference.
- Temporal linearisation: the ability of a diegetic sound to impose a sense of sequential real-time on a sequence of images that can be read as simultaneous or successive.
- Unification: providing atmosphere and bridging breaks.
- Punctuation: either sound or music can provide support for structure and clarity.
- Anticipation: convergence/divergence. Sound that have "... tendencies, they indicate directions, they follow patterns of change and repetition that create in the spectator a sense of hope, expectation, and plenitude to be broken or emptiness to be filled" (ibid. p. 55).
- Point of audition: attributing a direction to what is heard.
- Definition/fidelity: the distinction found between the acuity and precision in rendering detail and faithfulness to the true experience of hearing the sound.

- Phonogeny: the qualities in a voice that make it suitable for electronic reproduction. Sometimes at odds with a voice deemed 'nice' under normal conditions.
- Grain and unevenness: "...the impression of realism is often tied to a feeling of discomfort, of an uneven signal, of interference and microphone noise, etc." (Ibid. p.108).
- Materialising sound indices: "the materialising indices are the sound's details that cause us to "feel" the material conditions of the sound source, and refer to the concrete process of the sound's production." (Ibid. p.114).

Chion's achievement is not in inventing the sound objects that constitute the palette for the film soundtrack, but in identifying and, in many cases, naming them for the first time. Some, such as 'grain and unevenness', could be applied to synthetic voices straight away with little difficulty; others, such as 'temporal linearisation', are tied to a more subtle decision-making process, requiring aesthetic judgments that make it difficult to envisage a practical implementation in synthetic speech.

Chion's most relevant and impressive theory is 'the acousmètre', a conflation of 'acousmatic' (sound with no visual analogue) and 'être' (from the French verb 'to be').

One may argue that synthetic voices are *doubly* disembodied. In the conventional sense, much like characters in a radio play, they have no physical or visual embodiment, but unlike characters in a radio play, there is no embodiment of the *source* of the voice which, even at a theoretical level, can be made material. We cannot 'google image' them and thus imaginatively layer the voice with an image. Chion's 'acousmètre' posits a catalogue of properties we intuitively ascribe to a disembodied voice. In section 2.5.4, the particular power of a sourceless voice to influence and, sometimes, control those to which it is exposed was presented. One need only consider the denouement when the voice of the Wizard in the 'Wizard of Oz' (Baum, Garland, Haley et al. op. cit.) is finally materialised in unimpressive form to realise the latent power of withholding embodiment. The need to anchor the disembodied voice to a source was demonstrated in the device of the ventriloquist's dummy, but Chion's purpose is to catalogue the properties associated with a disembodied voice that may be used to advantage the drama, rather than to scare people. Chion's approach is to draw upon a wide range of examples from the cinematic oeuvre; these examples are not relevant to this

research. However, his analysis reveals a set of positions that may be held by a user when exposed to a disembodied synthetic voice; positions which will need consideration within the PAT framework.

- Does the user ascribe a source to the disembodied voice? If she does, is the source perceived as, for example, a character, a neutral nobody, a specific person or a characterless machine?
- Does the voice imply an environmental or geographical location (e.g. in the machine, on the screen, in a call centre, in the user's head)?
- What knowledge is a sourceless voice required to possess? (This point is made clearer if we consider a narrator in a play: does she always have complete knowledge of the drama, or can her knowledge be flawed?)
- Very simple parameters may modify the user's perception of the source of a disembodied voice: more resonance, for example, may suggest internal thoughts; more reverb may suggest projection or oration.

List 2: Principles for the user evaluation of a disembodied voice

In the previous sections we have reviewed many sources and perspective that provide insights into the interactive processes occurring in human and machine speech production and perception. From the problems of rendering stillness in robots through the preponderance of pauses in spontaneous speech to the use of emptiness to punctuate cinematic form, silence is a powerful signifier; but, as a signifier, it remains elusive. Decoding the information embedded in silence is a speculative art and the problem can only be addressed with some lapses into metaphysical means of expressions. Hence the title of the next section.

2.6 Metaphysical reflections on silence

“Silence undermines ‘bad speech’, by which I mean dissociated speech – speech dissociated from the body (and, therefore, from feeling), speech not organically informed by the sensuous presence and concrete particularity of the speaker, by the individual occasion for using language. Unmoored from the body, speech deteriorates. It becomes false, inane ignoble, weightless. Silence can inhibit or counteract this tendency, providing a kind of ballast, monitoring and even correcting language when it becomes inauthentic.” (Sontag 1969) cited by Darla M. Crispin (Losseff & Doctor 2007, p.138)

We can interpret Sontag’s position as an articulation of the potential for silence to counteract the awkwardness of the disembodied voice. This would miss the broader point that, from a modernist perspective, silence offers a sleek antidote to the clutter and confusion of contemporary noisy living. It may also be a response to the then (1960s) proliferation of transistor radios. At around the same time, John Cage was challenging the very definitions and distinction between sound, music and silence with compositions, writing and lectures which sought to materialise silence as an absolute. After a visit to an anechoic chamber he recognised his failure in this famous quote:

"When we think we hear silence, we are actually hearing many sounds. For instance, the traffic sounds outside the window. Or if we are in a chamber, we hear ourselves our blood flowing and our nervous system." (Cage 1973)

The tone of regret that such a seemingly-achievable absolute as true silence will always be beyond human experience may reveal why it constitutes such a significant part of many theological and spiritual customs (vows of silence, meditation and prayer). “What all silences have in common is their fecundity: they are pregnant with unanswerable questions.” (Losseff & Doctor op. cit. p.2). It may be that the metaphysical qualities of silence were better understood by the classical world that had two words for time: ‘chromos’, which could be quantitatively represented as the linear passing of time, while ‘kairos’ was the moment in-between which could only be defined subjectively, as a lack of ‘chromos’. Apply this principle to the temporal dimension of a musical score and “Silence can create an impression of atemporality which erases any sense of movement” (ibid. p.12). Resonances of the distinction

between the Newtonian concept of absolute time and Einstein's deeply metaphysical concept of relative time (at least for many, this is the only way of accessing Einstein's vision) also apply.

Silence is principally portrayed, even in a multidisciplinary context, as the negative absence of sound rather than the positive presence of silence. We should be reminded that "the idea of silence can be perceived as a form of communication – expressing reflection, for instance, as it might in Japanese culture" (ibid. p.1). Silence can present uncertainty, but it can also direct the listener to expect something. A question that remains deliberately unanswered may imply a threat to the poser of the question that will be resolved in a subsequent exchange. While we intuitively respond to silences in speech, we do not usually notice them. This is less the case in music.

Silence is deeply embedded in musical thought and training. Unlike actors, musicians may receive instruction for the location and duration of silences, either directly as written symbols in the score or indirectly from a coach, but judgments about the location and duration of silences is a significant part of the performance aesthetic for both actors (Fónagy 2001, p.598) and musicians. This is illustrated in a range of different musical styles on CD (Track 34), (Track 35) and (Track 36).

John Potter's view, expressed as "The Communicative Rest", is informed by his experience in live musical performance and is so rich in relevant points that it is worth quoting in full.

"Communicative silence is in large part about timing and is often a subtle combination of the predictive and the reflective. It is (ironically) often the point at which the performer and listener are closest to each other in intent, and the audience attention is at its most engaged. We tend to identify the delivery of a piece of music with its notes, and it is perhaps in part because of this that our attention increases when there is an absence of notes. The reflective element is primarily for the benefit of the listener: it has the capacity to enable him or her to make sense of what has gone before, to enter into the creative process of reconstructing the performer's meaning (or indeed of inventing it from scratch). There can be only nanoseconds of time involved and what the listener understands during that period may only be a kind of meta-thought, but at the very least there should be the illusion of something being communicated and the sense that something in the narrative may be about to change,

or may have already changed (illusion because performers cannot be sure that they are communicating anything at all: they can express a thought and hope that the listener will be able to make something of it). The predictiveness is fuzzy: the performer may well seek to narrow the interpretative options that may occur to the listener, or might wish to mislead for effect; the listener may have time to question what might be about to happen. For the performer it is a chance to vary the balance between predictability and unpredictability, and getting this equation right is the key to keeping the attention of the audience” (Potter J. in Losseff & Doctor op. cit. p.156).

Varying the balance between predictability and unpredictability and getting this equation right is also key to some of the PAT framework tests documented herein.

Potter makes points that would only be apparent to a performance practitioner. Silences that provide a point of contact with the audience in the decision-making process are critical in voice acting no less than musical performance. But the underlying unreliability of these moments to accurately communicate the intended meaning seems to trouble Potter. Not so our synthetic actors. We are after the point of contact and nothing more. If the point of contact is made then we are on the way to liveness. Potter also correctly highlights a key feature of silence: to offer a chance to balance predictability and unpredictability. This seems to correlate with the humanness-algorithms (designed to create controlled unpredictability), and the use of randomness to render them. Potter would rightly argue for the need for an intelligent (in his case, musically intelligent) entity to direct the decision-making process; however, given the difficulty for any entity, however intelligent, to wholly communicate what they intend to, this seems to confirm Widmer’s argument that the illusion of ‘expressive playing’ (see section 2.5.7) may not be difficult for computers. Silence may be an important contributor to this illusion

In speech synthesis, the usefulness of silence has had little attention. This seems a perfectly reasonable position when one considers the difficulties in representing its converse, and the relatively trivial nature of the rendering problem silence presents. Another reason why silence may have had little attention from speech researchers is that, just as in music, the information it carries is much less precise than the sounds it punctuates. Although it is possible to hypothesise about the semantics of silence (how silence is used to help create meaning in

speech), what the speaker intends by the silence cannot be reliably decoded. A long silence can mean (among other things):

- “I have said something important; think about it.”
- “I am about to say something important.”
- “I am so overwhelmed by emotion, I cannot continue.”
- “I have said something funny so I will wait for you to laugh.”
- “I have said something difficult so I will wait for you to understand.”
- “It’s your turn to speak.”
- “I have no idea what to say next.”

Probably with the exception of the last point, these silences are used to great effect by many politicians, among them the President of the United States, Barack Obama.

The imprecise meaning of silences would be a problem if our intent was to support or enhance the semantic acuity of synthetic voices. However, this is not the intent, as ‘liveness’ is unconcerned with the meaning of a text. It seems possible that long silences in a synthetic voice would carry some of the affect that silence carries in the examples above, and thus silence could provide the user with an anchor for a cognitive embodiment. The embodiment would not need to be anyone or anything in particular; it need only be an embodiment with a capacity to appear to pause for a purpose. Ideally, the purpose projected must be a positive one, and not, in the case of a synthetic voice, “I am silent because I am broken.”

Silence proved to be a significant contributor to the emergence of codifiable variables implemented in the PAT framework. The last sections of this chapter draw together the sources and perspectives to present a theoretical basis for the development of the PAT software tool and the PAT framework.

2.7 A theoretical basis for the PAT tool

The PAT software tool was required to test features of the PAT framework. The PAT tool had to operate within a pragmatic framework of available data and usable metrics. The potential to derive codifiable features for each of the perspectives discussed were considered. Table 5 represents the advantages and disadvantages of each feature presented by the perspective. As

silence has contributed to a number of the listed perspectives it is not considered in its own right.

Section	Perspective	Features	Advantages	Disadvantages
2.3	The 'Uncanny Valley'	Presents a goal for the optimal degree of anthropomorphism	Limits user expectations of the embodiment	May not be true for speech
2.4.1	Paralanguage	Sounds other than strict speech sounds included in human speech	Can be very expressive Many extant examples.	Some features difficult to encode An additive process May have semantic repercussions
2.4.1	Paralanguage	Pauses in spontaneous speech	Simple to encode and some metrics May help represent spontaneity which may help support liveness	May be perceived as errors
2.4.2	Paralanguage	Breaths and filled pauses	Simple to encode and some metrics.	An additive process. May be uncanny
2.5.1	Acting	A large repertoire of tried and tested techniques	Frames the interaction within a familiar context	None
2.5.3	Renaissance verse speaking	Pauses, breaths, end stopping, tone changes	Strong metrics. Some links with spontaneous speech.	Can be semantically derived Stylistically derived
2.5.4	Ventriloquism	Unusual vocal sounds highly characterised	Frames the interaction within a familiar and constrained context	May be uncanny
2.5.5	Singing and opera	Highly stylised techniques and sounds	Potentially very expressive	Not really speech May be 'over the top'
2.5.6	Extended vocal technique	Noises the human vocal tract can, but does not customarily make	Based on human vocal capabilities. Many extant examples.	Not widely accepted in current musical aesthetic May prove equally unacceptable when transferred to synthetic speech.
2.5.6	Expressive breath sounds	Use of breath to punctuate and inflect expression	May operate subliminally	May be uncanny
2.5.7	Musical expression	Plotting emotional valence (positive and negative emotions)	Well established aesthetic. Limitless examples.	Difficult to determine absolute metrics. Unpredictable subjective evaluation
2.5.8	Humanness algorithms	Controlled randomness.	Proven successes in musical applications.	Would require tweaking No evidence it works for speech
2.5.9	Liveness	A framework for evaluating the relative spontaneity of performance instances	Disambiguates real from live	Many variables with subjectively defined values
2.5.10	Cinematic sounds	Many techniques and treatments for sound and voice	Well tested DSP methods to achieve potentially rich and subtle results	May be closely aligned with visuals

Table 5: Perspectives and features to inform the theoretical basis for the PAT software tool

Some perspectives contained individual features that offered potential together with other features that appeared to offer less potential (or greater difficulties), thus the final list of features to be encoded in the PAT software tool included a selection from each perspective. Some features were derived from a combination of several perspectives. The derivation and the type are also noted. The final list is shown in Table 6:

Feature	Derived from	Type
Random perturbations to the speech stream	Humanness algorithms	Subtractive temporal variations
Grammatical pauses	Renaissance verse-speaking Spontaneous human speech Silence	Subtractive temporal variations
Non-grammatical pauses	Humanness-algorithms Spontaneous human speech Silence	Subtractive temporal variations
Periodic speech rate variations	Renaissance verse-speaking Spontaneous human speech Humanness-algorithms	Subtractive/distractive temporal variations
Breath synthesis	Renaissance verse-speaking Spontaneous human speech Extended vocal technique	Additive paralinguistic variations
Background sounds	Paralanguage Cinematic sounds	Additive paralinguistic variations

Table 6: Features encoded in the PAT software tool

2.8 A theoretical basis for the PAT framework

Unlike the PAT tool, the PAT framework is an abstraction. Accordingly, the framework is presented as a high-level model representing an idealised interactive environment between a human user and a synthetic voice. Also shown is the emergence of the framework from an imagined existing framework although, as discussed, none exists. Applications of the PAT framework are likely to dip in-and-out of the full range of sources and perspectives presented in this chapter. Some theoretical applications that dip in and out are discussed in (Chapter 8 Potential applications). The degree to which the framework is successful in representing this

complex collage coherently is the subject of subsequent chapters. For now the framework is presented diagrammatically.

The framework attempts to merge low-level prosodic manipulations with high level guidance on content and context, rendered through voice-acting, scripting and setting. The framework was tested in a number of scientific- and performance-based experiments and evaluations, culminating in an hour-long theatrical performance.

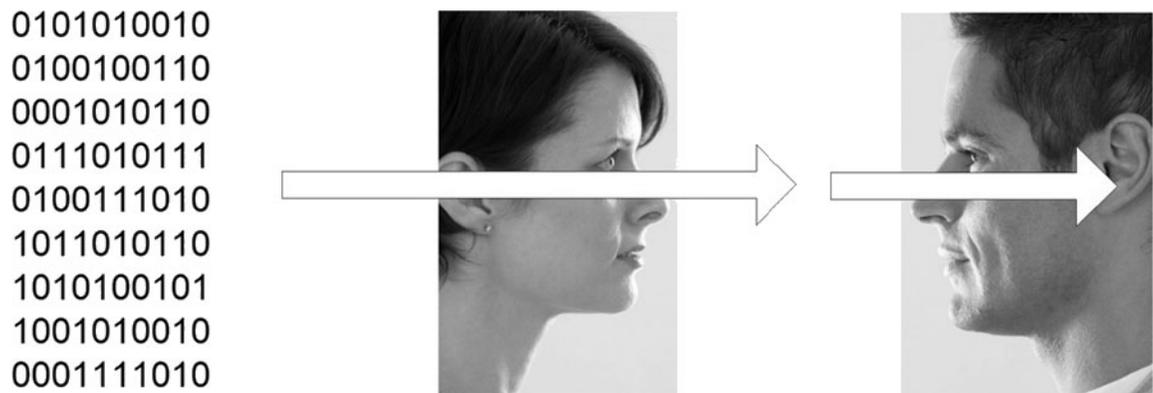


Figure 11: An aspirational synthetic speech system

In Figure 11, unconstrained computer code (code capable of capturing all the features that constitute realistic human-speech) finds a perfect unconstrained embodiment in a human-like artefact and the output is perceived as human speech by an archetypal, similarly unconstrained, listener. 'Unconstrained' means entities that are able to operate in a theoretical environment that either (a) presents no limitations or (b) can be said to present a normalised or neutral version of both speaker and listener. Neither (a) nor (b) are computable.

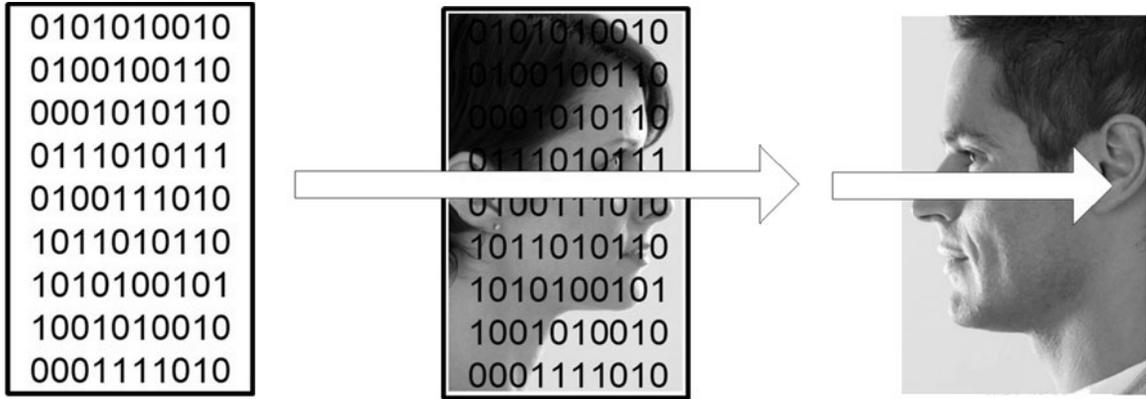


Figure 12: Current speech system outcomes

What seems to actually happen is shown in Figure 12. The limitations of the computer code produce a leaky and constrained embodiment of a human-like artefact that is perceived as false by the unconstrained user. (Constraints are shown by the addition of the black rectangle around the images). Neither the speaker nor the listener can be normalised in the way the designers would hope.



Figure 13: An alternative speech system

In Figure 13, the limitations of the computer code are accepted because the embodiment is framed as a theatrical performance and not a real human. The user expectation is appropriately constrained and falseness is accepted.

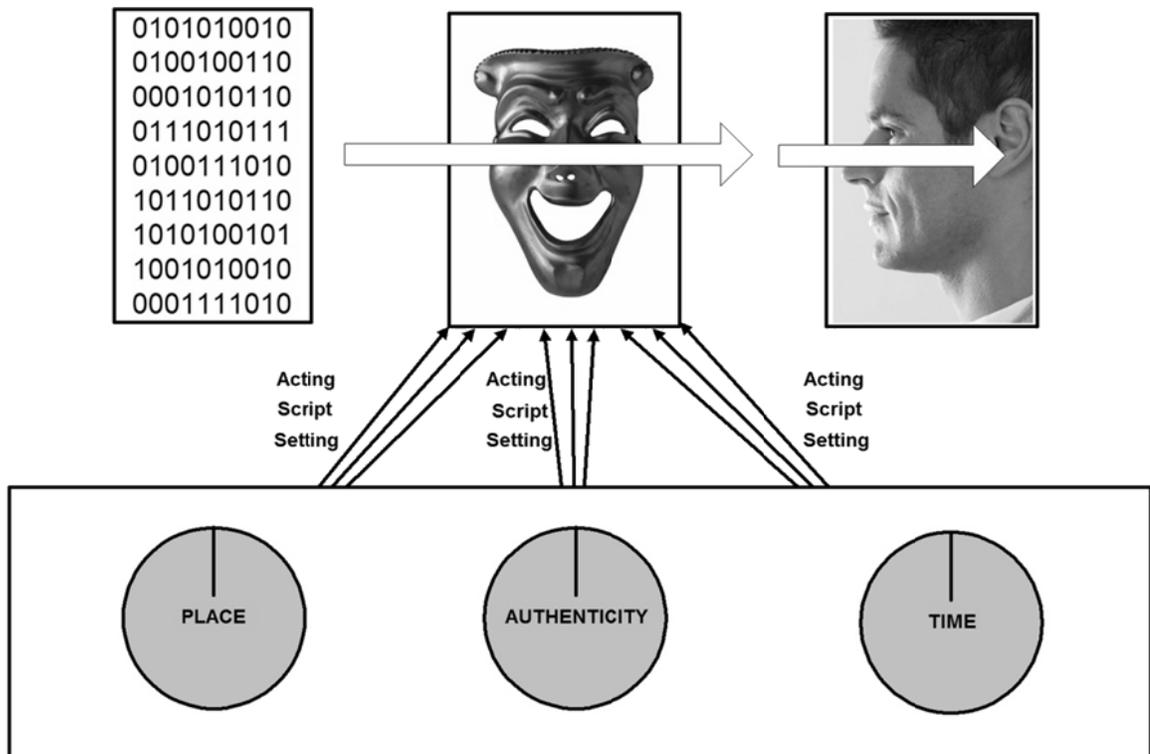


Figure 14: A “visualisation” of the PAT framework

In Figure 14, the PAT framework operates on the speech embodiment by manipulating three dimensions: place, authenticity and time. To recap:

- Place: locating the voice and the user in a shared auditory and physical space (fictional or actual).
- Authenticity: manipulating the user’s sense that the voice is being transparent or truthful.
- Time: manipulating the user’s perception that they are operating in a time frame shared with the voice.

The modifications are rendered using scripting, setting, and acting.

- The script is the actual words spoken by the voice, provided by the script writer.
- The setting is the location of the voice agreed with the user (this can be fictional or actual).

- The acting is the voice itself, and could include all the modifications available to the specific voice; modifications that may include voice style, gender, voice quality and all prosodic variables (similar to the modifications a human actor would have available).

2.9 Conclusions

In this chapter, the principle theoretical strands that support the propositions to be examined in this thesis and the emergence of the PAT framework are introduced. The theories are derived from a wide range of literatures, some of which resist a reductionist approach, giving rise to frameworks for the framework, rather than usable metrics. Acting, performance, abstract art, solving the Turing Test, ventriloquism and cinema all suggest perspectives that derive their power from the WSOD they expect of the user. An analysis of the features they have in common suggests the following list:

1. The artefact does not make any concessions to realism.
2. The user is expected to engage with an abstraction and to flesh out the illusion for herself using WSOD.
3. The artefact may delight in demonstrating the mechanism behind the illusion when appropriate.
4. The artefact may ascribe a source to itself. This does not have to be true.
5. The artefact may ascribe a location to itself. This does not have to be true.

List 3: Common features derived from the sources and perspectives

These principles provide the framework that supports the design of the PAT framework.

A pragmatic approach has been taken in determining the features encoded in the PAT tool by selecting six features (random perturbations to the speech stream, grammatical pauses, non-grammatical pauses, periodic speech-rate variations, breath synthesis, background sounds) from a much longer list. The other items on the list are held over for consideration in 'Conclusion and future research'.

The high-level theoretical basis for the PAT framework has been presented in diagrammatic form, and the render methods applicable to each PAT dimension have been explained.

3 The language of speech related phenomena

This chapter sets out to define a structured vocabulary, derived from the relevant sources and perspectives discussed in Chapter 2, to provide a clear interpretation of the key comparative terms used in this document. The PAT framework circumvents the use of terms such as 'life-like', 'natural' and 'realistic', preferring a non-anthropomorphic approach. However, these broad concepts provide comparative benchmarks for the evaluation of auditory properties exemplified by extant synthesis solutions and require analysis. In addition, a short discussion is presented exploring the difference between speaking and reading as it may relate to synthetic speech.

The language used and the frameworks envisioned in some of the more recent developments in HCI (referred to in the preface) suggest that significant synergies may exist with the other disciplines, such as the arts and humanities, where ambiguous problems are presented and less-rigorously-testable solutions have to be found. This has led to the emergence of fields that have a strong interdisciplinary flavour and may indicate a need for agreed interdisciplinary lexicons.

More holistic criteria for evaluating HCI systems have become more commonplace as the usage of some computer systems has shifted from goal-orientated experts to entertainment-orientated consumers. Usability, measured in conventional HCI terms, may not be enough for a system or product to succeed in markets where evaluations based on experience, fun, affect, life-style and fashion predominate. Implementations of synthetic speech operate in these environments, and, accordingly, ways of accurately describing the properties of these new domains are required. As Eide et al. say

“We can well imagine that when a variety of TTS voices hits the market, all equally good in terms of quantitative measures such as perceptual bandwidth and cognitive load, some will nevertheless become more popular than others. Designers of new voices will then attempt to add novel twists to the currently most popular ones and

some of those will again become standard. In this way “fashion” in TTS voices will undoubtedly keep evolving, somewhat resembling a random walk.” (Eide, Bakis, Hamza et al. 2005)

Many of the HCI theories that have emerged from this new generation of requirements are evolving now. Some, such as ‘Funology’ (Blythe 2003), rest upon a relatively small body of literature, while others, such as ‘Computers as Theatre’ (Laurel 1991), rest upon a substantial part of the Western canon. The problem that remains in all cases is communicating accurately with a mixed audience who may hold radically different definitions of key terms.

For example: affective computing (Picard 1997) anticipates a human-machine relationship based upon recognizing and responding to emotional states. Robust models of emotion drawn from the psychological sciences remain elusive, and the language of emotional states negotiated by the theory remains as grounded in the arts as in the sciences. ‘Technology as experience’ (McCarthy & Wright 2004) describes HCI that can be measured according to metrics derived from day-to-day pleasurable and aesthetic experiences, or ‘felt life’. This is an overtly holistic concept with deliberately ill-defined borders between the system and the variables that impact upon the user’s evaluation of the system; variables found in the mundane experiences of everyday life. In ‘Computers as Theatre’ (Laurel 1991), Laurel argues for a deeper appreciation of the synergies that can exist between theatre and HCI. She draws upon Aristolean Poetics (Aristotle, Roberts, Bywater et al. 1954) to model the human computer experience similarly to the way a dramatist might model the audience’s experience of a play. Mateas investigates the notion of “artificial Intelligence as art”, and creates a computer game, ‘Façade’, in which the characters are imbued with deep and subtle emotions, similar in complexity to those expressed by characters in a television soap opera (Mateas 2002). Alan F. Newell³³ (2006) uses techniques derived from the performance theories of Boal (Boal 2002), working in conjunction with a theatre director and team of actors to support the design of HCI systems for older users. Also of interest is research into intelligent anthropomorphic embodiments, or life-like characters (Prendinger & Ishizuka 2004, Rist & Andre 2000, Rizzo 2000, Trappl & Petta 1997). These artefacts take on human-like roles, like synthetic news readers, characters embedded in interactive toys, helper robots and web-based virtual actors. They are required to show, among other things, emotion, personality, empathy and character.

³³ No relation to the author.

Although each of these new approaches may be labeled interdisciplinary, the work tends to have difficulty achieving a broad distribution of knowledge across disciplines: for example, all of the examples listed above would probably only be well known within the context of computer science, despite their relevance to other disciplines. Even within the same subset of a discipline, it is not easy to tease out how individual authors discriminate between the uses of their key terms. This problem as it relates to the main theme of this research is illustrated in the next section.

3.1 Illustrating the problem of ambiguous terminology

In the field of synthetic characters, the definitions of lifelike (very relevant to the codification of liveness set out herein) include:

- the seven qualities of life-like characters posited by Hayes-Roth: “conversational, intelligent, individual, social, empathic, variable and coherent” (Hayes-Roth op. cit. p.447).
- “emotional and believable agents” (Prendinger & Ishizuka p.4 citing Bates).
- “personality” as a key constituent of lifelikeness (Trappl & Petta pp.1-3)

It is interesting to note that there are no terms common to all three examples.

Other examples may be found in the field of speech synthesis where the term ‘naturalness’ may be interpreted as:

- “A function of much improved prosodic modeling” (Keller 2002, p.4).
- “A waveform which a human listener judges unambiguously to be a speech signal produced by a human being” (Tatham & Morton 2004, p.83).
- “How well it reflects the speaker’s emotions and/or how well it features the culturally shared vocal prototypes of emotions” (Keller op. cit. p.237).

Thus, within two areas relevant to this research, consensus in the use of specific comparative terms cannot be assumed. There are too many contentious terms prevalent in the literature to attempt to define all of them, but without some standardisation of a select vocabulary, the

objective evaluation of new synthetic speech solutions that may emerge from interdisciplinary practice is likely to be impaired.

The solution proposed is to rationalise the number of terms used in comparative evaluation and simplify the meanings. A further refinement is to apply the principle of differential inheritance to the lexicon, obliging all the properties of the parent term to be inherited by the child without the new properties of the child term becoming present in the parent. In the next section, a strategy to select the key comparative terms for use in this thesis is presented.

3.2 Addressing the problem of ambiguous terminology

In order to address the problem of ambiguous terminology, a structured vocabulary of comparative terms will be utilised in this document. To arrive at the list of comparative terms, a review of the relevant literatures was conducted, looking for common comparative terms that may have uncommon usages. The selection was further refined on the basis of usefulness within this thesis (although not all the terms are used repeatedly); therefore, the term had to be one that could feasibly be used to describe the auditory properties of a synthetic voice. In each case, the subtleties of meaning are eliminated and replaced by a composite term which takes on a delimited interpretation. The definition is not intended to capture any aspect or to amalgamate existing definitions; rather it presents a new one which may actually contradict current usage in one or the other field. An example would be the term ‘natural’ as defined by Keller: “how well it reflects the speaker’s emotions and/or how well it features the culturally shared vocal prototypes of emotions”. In the structured vocabulary ‘natural’ is a parent of ‘emotional’, and cannot have the property ‘emotional’, thus Keller’s definition would be invalid.

Other key terms specific to the synthesis literature that provide crucial components for the structured vocabulary are also included. What emerges is a structured vocabulary for the evaluation of a speech artefact, applicable to this thesis, which could be freely adopted in other speech evaluation environments.

3.2.1 The structured vocabulary

The structured vocabulary presents a six point scale of comparative or evaluative terms that will be used in this thesis and could be used more generally to compare synthetic voices. 'Affective voice' is the highest attainment in synthetic speech production (level 6) and 'speech' the lowest (level 1). All human speech is positioned at level 6, and the synthetic speech examples presented in this research achieve a minimum of level 2.

The PAT framework is designed to modify speech at level 2 so that it is perceived by users as if it were speech at levels 3 or 4.

The structured vocabulary in Table 7 illustrates the key terms derived from the literature, the composite term used to capture several key concepts, and a brief description of the proposed usage of the term.

Terms from the literature	Term used in this thesis and level	Proposed usage
Neutral speech	1. Neutral speech	Neutral speech does not exist either in speech or synthetic speech although the meaning intended by the term is speech that lacks all the features listed in the rows below.
Speech	2. Speech	An auditory signal that contains sounds that may be decoded as words, phrases and other units designed to communicate lexical information to the recipient.
Natural, Believable, Realistic	3. Standard speech	Exhibiting auditory features that are broadly intended to represent those found in groups of normal speakers.
Voice	4. Voice	Exhibiting auditory features that are intended to represent those found in standard speech in a single individual, however in synthesis this may be extended to include a group of speakers with the same voice.
Personality, Character	5. Character voice	Exhibiting auditory features that are intended to represent those found in non-standard speech that can only exist in a single individual.
Expressive Emotional Mood	6. Affective voice	Exhibiting auditory features that are intended to show modifications subject to the state of mind of the speaker

Table 7: Structured vocabulary for evaluating synthetic speech

3.3 The key speech terms for this thesis

In this section clarifications of the definitions applied to the terms used in this thesis is provided based on the literature reviewed in chapter 2. In each case the definition as presented in the structured vocabulary is displayed for easy reference. Readers are reminded that in this structured vocabulary inheritance is cumulative i.e. an affective voice³⁴ level (6) has all the properties of the voices in the parent levels plus the new property (in this case affect). First instances of each of these terms in the ensuing chapter of the thesis will cross reference to this section as a reminder to readers that the terms are to be used with precisely delimited meanings.

3.3.1 Neutral Speech

Neutral speech does not exist either in human speech or synthetic speech although the meaning intended by the term is speech that lacks all the features listed in the ensuing definitions.

In this thesis, 'neutral speech' will be taken as a theoretical concept with no equivalence in the human speech-production system. It is analogous to the 'core' of sound prior to modification by the human vocal tract.

It is the lowest-order feature in this structured vocabulary.

3.3.2 Speech

An auditory signal that contains speech sounds that may be decoded as words, phrases and other units designed to communicate linguistic information to the recipient.

'Speech', as used in this thesis, is a lower-level property than 'standard speech', as it represents the raw data prior to encoding as 'standard speech'. 'Speech' can be heard in some

³⁴ Logically the only speech that can be rendered by the human vocal tract is affective speech. None of the lower order speech types can ever be heard and exist only as convenient theoretical labels. Even the speech output by synthetic speech systems can only be perceived as affective speech, despite having no emotion present. It is the quality of rendering and appropriateness of the affect that is in question.

of the very earliest examples of synthesis, where the listener would struggle to comprehend the intended communication, despite the fact that the information was represented in some form. Modern synthesis systems have achieved 'standard speech'.

3.3.3 Standard speech

Exhibiting auditory features that are broadly intended to represent those found in groups of human speakers.

'Standard speech', in this thesis, represents a kind of comfort zone. In 'standard speech', nothing offensive or startling gets through, but equally nothing that could be taken as individual passes either. In 'standard speech' intelligibility is critical and all other factors are sublimated to this goal. 'Standard speech' is wholly transferrable, with multiple instances of standard speakers allowed to exist.

3.3.4 Voice

Exhibiting auditory features that are intended to represent those found in standard speech in a single individual; however, in synthesis, this may be extended to include a group of speakers with the same voice.

The distinction between 'speech' and 'voice' is the most troubling issue included in this short review, with little consensus across disciplines.

For example Elsam suggests that 'Voice is quite simply the sound you make. Speech is the use of lips and tongue to shape and to control that sound into words' (Elsam 2006, p.98) while Trask reverses the position and defines 'voice' as "The natural and distinctive tone of the speech sounds produced by a particular person" (Trask 1996, p.378).

One way of capturing the distinction is to focus on the role played by an individual speaker in spontaneously modulating vocal sounds and call the resulting output 'voice'. (Caelen-Haumont 2002, p.353) describes the difficulty in representing spontaneous speech as '...variability, adaptation to communication context and addressees, and ultimately, subjectively identified speech characteristics at every level from acoustics to semantics.' Thus Caelen-Haumont seems to identify the rich variation in voices as an outcome of complexities resulting from the

production of spontaneous speech. This distinction between ‘speech’, ‘spontaneous speech’ and ‘voice’ is significant in appraising the relative complexity and difficulty of providing a computer representation. It could be argued that computer ‘speech’ has been realised, and that the current research agenda is directed to the production of Caelen-Haumont’s definition of ‘spontaneous speech’, with all its incumbent subjectivity and complexity. Because the human voice production system is a non-linear system, a given input does not produce a predictable output, and it is sensitive to external factors (such as environment, context, addressee) as well as internal variables, including mood, emotion, pragmatic and linguistic demands, errors and paralinguistic factors. ‘Voice’ synthesis, as distinct from ‘speech’ synthesis, may be incomputable, at least with current technology. ‘Voice’, in this thesis, will always be presented as a higher-level feature than ‘speech’, requiring more complex rendering techniques.

3.3.5 A character voice

Exhibiting auditory features that are intended to represent those found in non-standard human speech that can only exist in a single individual.

A ‘character voice’ is a lower-level property of voice than an ‘affective voice’. A ‘character voice’ may present affect but the degree of nuance and subtlety expected is less. A ‘character voice’ is analogous with a cartoon voice-over, which may rely on exaggeration and pastiche to create instantly recognisable types rather than rounded portraits of real human-beings. However these types are not fully interchangeable. Bugs Bunny may be an exaggeration but he is unique. In rendering characters of this type, the voice ‘impersonates’ rather than ‘acts’. Defining the term in this way aligns it with the contemporary disdain for ‘character’ in theatre and performance, where the term has acquired a pejorative meaning, as well as the lower status afforded to an impersonator than an actor.

Producing ‘character voices’ is not a goal of this research, and the term will only be used in reference to the work of others.

3.3.6 An affective voice

Exhibiting auditory features that are intended to show modifications subject to the state-of-mind of the speaker.

‘Affect’, in speech, relates to the emotional content encoded into the speech stream. As such, it is a high-level property of voice derived from physiological manifestations of feelings, themselves derived from physiological responses to various stimuli. A significant amount of work has been done in this area. (Burkhardt 2009) provides an excellent collection of recorded examples, from the pioneering work by (Murray 2009) to more recent examples. ‘Affective’ synthetic voices are already in the mainstream of synthesis research, if ‘irritation’ is considered a worthwhile affect. More likely, the term ‘affective voice’ is meant to indicate a voice capable of demonstrating a range of subtle, nuanced emotions, just as the human voice does. The representation of ‘black-and-white’ emotions such as ‘hate’ or ‘fear’, and moods such as ‘friendly’ or ‘formal’, has already achieved a commercial presence (Loquendo 2008), but automating the generation of appropriate and convincing emotional speech is some way off. The goal of ‘affective’ synthetic speech is part of a broader initiative directed towards ‘affective computing’, in which the computer demonstrates sensitivity to emotional input as well as output. This is proving very challenging.

One reason for the difficulties can be easily demonstrated using the board game ‘Moods’³⁵ see Figure 15. In the game the players draw pairs of cards. On one card will be a phrase (e.g. “Frankly, my dear, I don’t give a damn.”) on the other card will be a mood (e.g. “sneaky”). The objective is to say the phrase communicating the mood, in order that the other players guess the mood. A well-practiced actor may be quite accurate at conveying the appropriate ‘mood’, but most players will find it a very difficult task. ‘Mood’ is only one of the many variables requiring computation to produce a convincing affective voice; and, considering anecdotal evidence from the game, the fact that most *humans* have difficulty feigning ‘mood’ suggests that computers will find it near-impossible.

³⁵ © 2001 Hasbro International Inc.

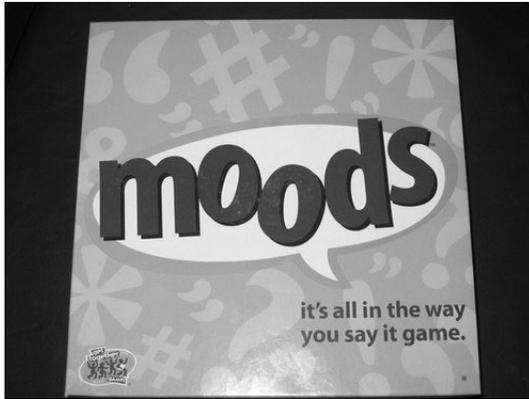


Figure 15: The game 'Moods' (©Hasbro Inc)

Therefore, we will consider an 'affective voice' as higher-level synthesis; and such a voice is not the goal of this research. The term, therefore, will only be used in reference to the work of others.

3.3.7 Speaking and reading: a discussion

In this section, the distinction between 'speaking' and 'reading' is considered. Although 'reading' is not conventionally a comparative term, and therefore is not included in the structured vocabulary, the first instance of the term 'reading' in the thesis will refer back to this section to remind the reader of the delimited meaning intended.

'Written' and 'spoken' texts are completely different. 'Written' text is structured in complete sentences, constructed so as to facilitate reading or comprehension. While we may think that we also talk in complete sentences, any examination of transcribed speech will quickly demonstrate that this is not the case (Goldman Eisler op. cit.). It is thus usually quite easy for a listener to distinguish between a piece of text being read and a person speaking spontaneously. Public speakers are generally more engaging if they do not read their text.

Of course, there are anomalies. An actor is presenting a written text, but has to do so in a manner which does not sound like reading: "The purpose of voice acting is to get 'off the page' with your message. Make it real..." (Alburger 2002, p.1). In this case, the actor will be assisted by the writer who will have written the dialogue to resemble spontaneous speech. Yet even a public speaker who has to read for one reason or another will try to disguise the fact by varying the inflection, looking up at the audience, adding impromptu remarks, or dramatically pausing.

One fundamental question to ask: 'is a TTS system "speaking" or "reading", and thus which term best describes the goal of this research? Superficially, the answer would be 'reading', in that a piece of written text is given to the machine, and it says the words which are written. However, a better analogy might be that of giving the 'script' to an actor, who will then attempt to read it in such a manner that it sounds spontaneous (has 'liveness')³⁶. In the studies documented herein, defining liveness to participants has been problematic. One solution adopted has been to use the distinction between 'speaking' and 'reading' to direct the listener. It may also be interesting to note that as a shorthand to encourage an actor to imbue a performance with more 'liveness', the distinction can particularly apt (see 4.2). Instances of reading in this thesis, and in the context of voice-acting, can be said to stand as the opposite to liveness.

3.4 Conclusions

This chapter has attempted to address the extremely knotty problem of defining precise comparative terminology to be used when comparing synthetic voices. The terms selected for analysis are those likely to be applied in an evaluative context to instances of synthetic speech. They are:

1. Neutral Speech
2. Speech
3. Standard speech
4. Voice
5. Character voice
6. Affective voice

By referencing the structured vocabulary the reader is able to confirm the precise meaning intended in this thesis by commonplace comparative terms. In addition to this solution,

³⁶ This concept is slightly more complex in execution than it may appear. This is illustrated on CD Track 37, in which John Humphrys (a very well-known radio interviewer, presenter and journalist) has been required to act the role of 'himself' in a radio play. Listeners familiar with Humphrys's manner of speaking will detect the effort he has made to sound more spontaneous than is required in his normal role. This is despite the fact that the character he is playing (himself) and the context (a radio reporter) requires no additional spontaneity, liveness or acting.

potential problems that may exist in interdisciplinary research which relate to shared lexicons and ontologies have been reviewed. A short discussion on the distinction between speaking and reading and the relevance of these terms to identifying instances of liveness concluded the chapter. In the next section two studies are documented designed to tease out the meaning of critical terms by eliciting user evaluations.

4 Detecting and describing liveliness and liveness

This chapter examines the problem of defining 'liveness'. A number of different terms were considered for the dependent variable, prior to the eventual selection of 'liveness'. The term 'liveliness' was used in the first two studies reported in this research. To avoid confusion with subsequent tests in which other terms and derivatives of 'liveness' are used, these two 'liveliness' studies are given a short chapter of their own. This chapter documents the two studies.

Despite the efforts documented in the last chapter to define a structured vocabulary of terms specifying evaluative criteria for synthetic speech, choosing a reliable single term, capturing both the intuitive sense of improvement and the disassociation from humanness, was very difficult. The term would need to be one that users could understand, detect and grade.

In the first studies, two terms were considered as appropriate descriptors for the quality desired in synthetic speech: 'spontaneity' and 'liveliness'. In discussion with colleagues, 'spontaneity' was rejected on the basis that due to certain predilections against synthetic speech, users would probably be unprepared ever to report the speech output from a computer as 'spontaneous'³⁷. Arguably, it is also an absolute term with no gradation possible; something is either spontaneous or not.

'Liveliness' is equated by Gustafson and House with 'fun' (Gustafson & House 2002). 'Fun' was likely to be easier to detect and reported more accurately by users in the context of synthetic speech. 'Lively' is also a term in common usage in many domains and discourses - "He has a 'lively' character," or "The pub had a 'lively' atmosphere," or "The debate was 'lively.'" are all examples of commonplace usage. Metrics for 'lively' exist in musical notation. In its Italian form, 'lively' - 'allegro' - defines a speed variously reported as between 120 and 168 beats per minute (bpm), and, accordingly, in music, 'lively' has a generally high emotional valence or positive association, which, it may be argued, is also the case with 'liveness'. 'Liveness' earns

³⁷ Despite this, 'spontaneity' is the term that maps most neatly to the concept of 'liveness' when the predilection against such a capability in synthetic speech is put aside.

its positive associations through the notion of authenticity or 'nearer to alive' and this may also be the case with 'liveliness'.

The term 'liveliness' was subject to the studies described in the next two sections.

4.1 Liveliness terminology test

The goal of this study was to establish whether the term 'liveliness' had the potential to present a sufficiently precise meaning or set of meanings to be reliably detected and reported by users. The objective was to use the term as a qualitative measure for subsequent tests on synthetic voices.

Participants: The test was undertaken with twenty-five first-year undergraduate students who were studying a range of interdisciplinary arts and technology-based subjects. Twenty-four were in the 18-19 years-old age range, with one mature student of unspecified age. Some had involvement in the performing arts and some played musical instruments³⁸, hence a bias toward interpretations related to these domains may be present. The study was conducted at the end of a teaching session on a voluntary basis.

Method: Each student was provided with a blank sheet of paper and asked to write down anything that came into mind when asked to consider 'liveliness'. They were asked not to confer. They were given five minutes to consider and record their answers. The papers were collected by the researcher for analysis.

³⁸ Most of the studies documented in this thesis collect additional data on gender, whether the participant is an actor, the participant's ability to play a musical instrument and their experience with synthetic speech. This was done in order that any hypothetical future studies conducted to establish a relationship between musicianship/acting experience or gender specific traits and participant responses to synthetic speech could utilise this data. In addition, participants with significant experience of synthetic speech could confound results. Noting the inclusion of these questions is documented for completeness sake but this data has not been chosen for consideration in this research at this stage.

Concept expressed	Typical Wording
Passion	To deliver 'liveliness', the performer requires passion.
The type of experience	'Liveliness' is experienced in the totality of a live performance, not in the parts.
Spontaneity	'Liveliness' implies the freedom to change spontaneously.
Personal interpretation	'Liveliness' suggests a personal interpretation.
Evoking realism	'Liveliness' evokes a realistic happening.
Interaction	'Liveliness' implies an interactive involvement with the performance.
A real-time event	'Liveliness' can only be experienced in real-time.
Inspirational	'Liveliness' encourages a freedom to excel.
Unique	'Liveliness' cannot be repeated.
Errors	'Liveliness' implies the freedom to change and to make errors.
Indicating tempo and mood	'Liveliness' is upbeat, bubbly.
Synonyms	'Liveliness' is positive, animated, energetic, full of life, skittish.

Table 8: Conceptual groups defining liveliness

Results: Analysis was conducted on the basis of a simple keyword search followed by subsequent categorisation by the researcher. Table 8 shows the results of the test reformulated with indicative rewording of the original statements by the researcher. The column 'Concept expressed' categorises the subjects' actual words and phrases into the twelve main conceptual groups identified by the researcher. This table does not indicate the frequency of words or concepts.

This interpretation of the data details properties that this group of subjects would accept as indicators of liveliness. It should not be taken as a requirement to include them all, as many subjects responded with only one concept.

Conclusion: The results indicate a broad range of definitions within a diffuse conceptual field. This may be exacerbated by the requirement to write the definitions down, rather than to simply report an occurrence of liveliness (this is addressed in 4.2). It is possible that 'liveliness' is a property like 'colour': easy to detect but difficult to describe. A more detailed test and analysis may have exposed some of the concepts as outliers, but an informal interpretation by the researcher (Table 9) of the data shows the broader categories of properties.

Property	Expressed by
Individualism	Freedom to change, interaction, personal interpretation, cannot be repeated
Positive mood	Upbeat, bubbly, animated
Authenticity	Realism, real-time, unrepeatable, a totality

Table 9: Categories of properties for liveliness

As a requirement for improved synthetic speech, the term ‘liveliness’ is shown to have complex linguistically defined properties, some of which are relevant to the objectives of this research. However, this study also showed a preponderance of properties associated with positive mood and this may lead to evaluations based primarily on irrelevant factors that may project this emotion. For example, there is a danger that fast articulation (as indicated by the musical specification of ‘allegro’ for ‘liveliness’) could be generally acknowledged as an instance of ‘liveliness’ and the other more complex properties described in Table 9 may be subsumed. Any single term encapsulating the required properties will have deficiencies particularly when subject to rigorous lexical scrutiny. These deficiencies may be less problematic when the term is evaluated in a more realistic context: accordingly from this study it may be concluded: that for the term ‘liveliness’ to be a useful descriptor of a requirement for synthetic speech production, the ability of users to detect and accurately report occurrences of liveliness rendered aurally should be tested. This is the subject of the next section.

4.2 Hearing liveliness test

In the first part of this test, the objective was to determine if the participants could correctly distinguish a recording of a human reading a text in a deliberately ‘lively’ style from reading (see discussion 3.3.7) the same text in a neutral style (see structured vocabulary 3.3.1). A third style, ‘performed to a live audience’, was added to increase the discriminatory challenge of the test. The qualitative difference (in terms of the objectives of this research) between the ‘lively’ reading and the reading performed to a live audience was not significant; therefore, the objective was really to see if the read text could be consistently distinguished from the other two.

In the second part of the hearing liveliness test, the objective was to determine if the participants could correctly identify a recording of text performed in a deliberately 'lively' style from a read text, in both cases with musical tones substituting for the words. As in the first part of the test, the third style was added with the same objective. This was designed to reveal if liveliness could be reliably encoded and decoded aurally without any linguistic content. This could be helpful when making the transition from normal human speech to unusual and synthetic speech in subsequent studies. A lower level of accuracy than the first part of the test was anticipated.

Participants: The test was undertaken with eighteen undergraduate students studying a range of interdisciplinary arts and technology-based subjects. None of the participants had taken part in the 'liveness terminology' test. All were in their late teens or early twenties. Twelve (a high proportion) played musical instruments (see footnote³⁸). The study was conducted at the end of a teaching session on a voluntary basis.

Method for the first part of the test (Test 1): An experienced male voice actor was asked to record three readings of the same section of text using a digital recording device. The actor was left alone, with the researcher outside the room, to make recordings of readings 1 and 2, and placed in front of a small audience of approximately 20 persons for reading 3. Each reading method was explained to the actor after the conclusion of the previous and no prior discussion of the text or the experiment took place. A short level test was undertaken before each of the readings took place. The text chosen was the lyrics to the song 'Fitter Happier'³⁹ by English rock group Radiohead, from their record 'OK Computer' (Radiohead 1997) (see Appendix C for the full text). The choice of text offered some ambiguous, metaphorical content together with some more explicit language and obvious sentiment. Grammatically, the text is segmental, with very short phrases. The actor recorded readings of the complete text in 3 ways.

1. Sight-read (read without having seen the text before)
2. Read in a 'lively' style
3. Read to a live audience

A short section of speech (approx 17 seconds) was selected and an audio CD transfer was made of the three tracks. These can be heard as CD Track 38 to Track 40.

³⁹ This song is one of the few examples of an artwork produced using speech synthesis (see section 1.1.5)

Method for the second part of the test (Test 2): The three recordings specified above were converted to MIDI note values using Digital Ear (Epinosis 2009). Subsequently, each recording was re-recorded, played on a MIDI instrument emulating a saxophone in the original register of the actor. In the process, quantisation was applied to the velocity data, giving a uniform velocity value to all the notes to assist audibility, but the original pitch (in microtones) dynamics and rhythms were retained. The actor's prosody was retained largely intact (except dynamic contrast) while the linguistic content was removed. The same short section (approx 17 seconds) of the recording used in the first part of the test was selected and an audio CD transfer was made of the three tracks, CD Track 41 to Track 43.

Procedure: The eighteen participants were asked to complete an individual questionnaire. The instructions for the test were presented on the questionnaire; consequently, no discussion of the procedure took place. All the participants listened to the six tracks in two groups of three tracks played on a professional CD player over high quality hi-fidelity speakers in a university seminar room; first the three recordings of real speech and then the three recordings of speech converted into saxophone tones. The order in which the recordings were played to each group was randomised, but, by chance, in both cases, the lively performance came first. The participants were asked to identify which was which and enter their answers on an answer sheet after they had heard all the recordings in each set of three.

Caveats: It was theoretically possible for a participant to use their memory of the spoken version of the test to inform their choice in the tones version of the test, although this would have been very difficult to do. The order factor may have also played a part in facilitating comparative evaluations. There is some evidence that the musicians in the group had an advantage, as five of the twelve participants who played musical instruments correctly identified all the recordings, against one of the six non-musicians.

An example answer sheet is reproduced as Figure 16:

Your Name	
-----------	--

Do you play a musical instrument Yes or No	Yes
---	-----

TEST 1

You will hear 3 short speeches.
You are required to identify the speech where:

- 1. The actor is reading the text for the first time
- 2. The actor has been asked to read the speech in a lively way
- 3. The actor is performing the speech to a live audience

DO NOT FILL IN THE FORM UNTIL YOU HAVE HEARD ALL THREE

	Reading	Lively	Performing
Speech 1		✓	
Speech 2	✓		
Speech 3			✓

TEST 2

You will hear 3 short speech-like phrases without discernable words.
You are required to identify the phrase where:

- 1. The actor is reading the text for the first time
- 2. The actor has been asked to read the speech in a lively way
- 3. The actor is performing the speech to a live audience

DO NOT FILL IN THE FORM UNTIL YOU HAVE HEARD ALL THREE

	Reading	Lively	Performing
Phrase 1		✓	
Phrase 2			✓
Phrase 3	✓		

Made a measured judgement, would have liked to hear each clip twice.

36

Figure 16: Sample answer sheet for the hearing liveliness test

Results

Table 10 shows graphs for the frequency distribution for answers to each question. The correct answer is shown in the heading for the graph and highlighted in grey.

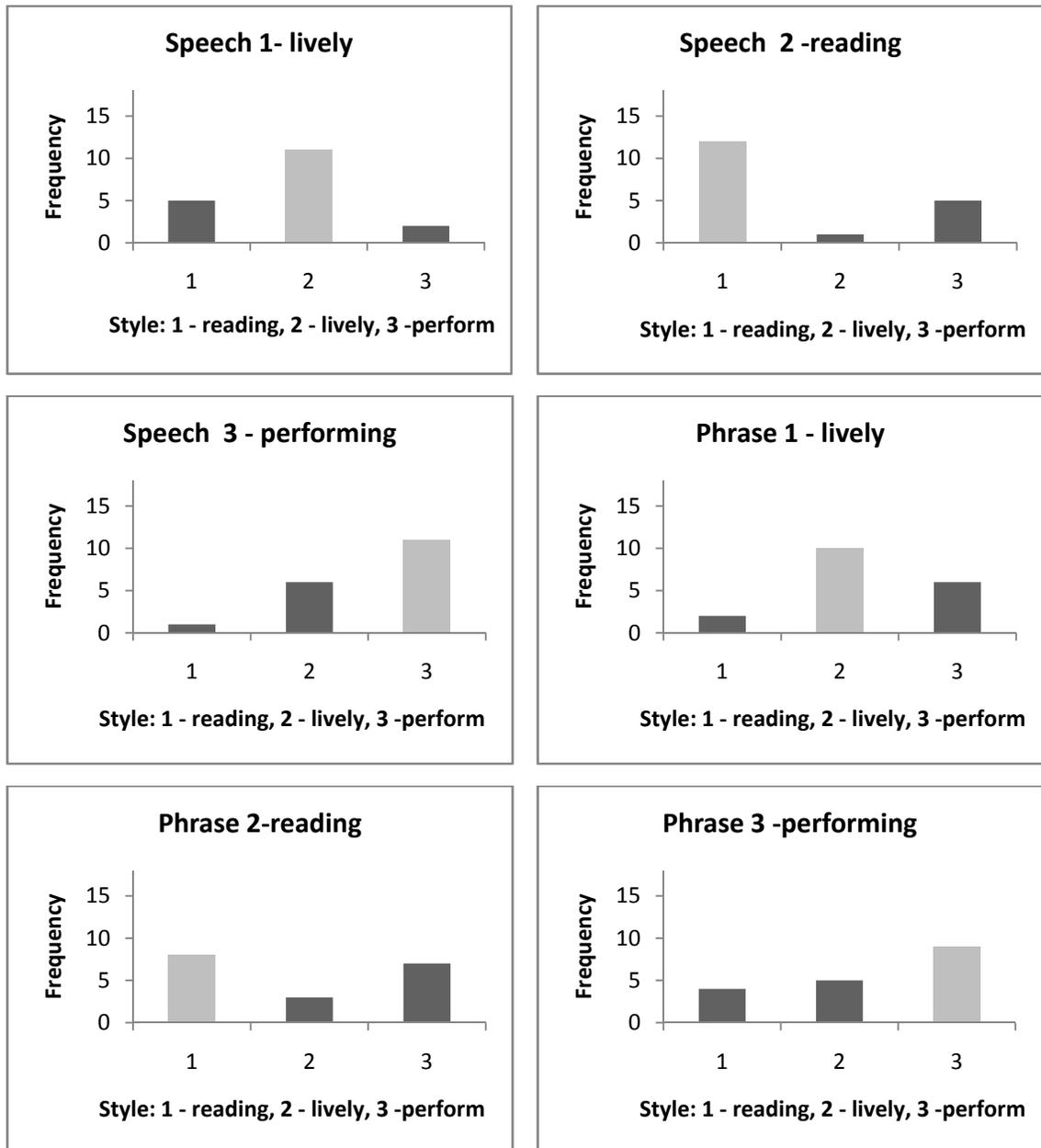


Table 10: Frequency for identification of recordings 1 to 3 and tone phrases 1 to 3

	Identified as lively	Identified as reading	Identified as performing	P =
1. Lively recording (speech)	11	5	2	0.030
2. Reading recording (speech)	1	12	5	0.006
3. Performing recording (speech)	6	1	11	0.016

Table 11: Results from test 1: hearing liveliness test – speech. The P column shows the p-value for a chi-square test of this data as reported by Excel software.

	Identified as lively	Identified as reading	Identified as performing	P =
1. Lively recording (tones)	10	6	2	0.069
2. Reading recording (tones)	3	8	7	0.311
3. Performing recording (tones)	5	4	9	0.311

Table 12: Results from test 2: hearing liveliness test – tones. The P column shows the p-value for a chi-square test of this data as reported by Excel software.

In Table 11 and Table 12, the results from test 1 and test 2 are shown.

The probability of the distribution occurring randomly was calculated using the χ^2 test in Excel.

For Test 1 the χ^2 test shows:

- Statistical significance ($p=0.030$) for the correct identification of the lively speech
- Statistical significance ($p=0.006$) for the correct identification of the read speech
- Statistical significance ($p=0.016$) for the correct identification of the performed speech

For Test 2 the χ^2 test in shows:

- No statistical significance in any of the tests

4.3 Conclusions

The results show that examples of lively, performed and read speech can be distinguished from each other. From this, it may be concluded that the three terms may be used in a comparative context as distinguishing descriptors for human speech.

Removal of the linguistic content results in none of the descriptors being reliably detected. The failure to show a conclusive result in relation to speech-like tones does not show that prosodic modulation (based on pitch and rhythm but without linguistic content) cannot communicate 'liveliness' but it may show that the underlying prosodic modulation has to be more extreme than that produced naturally by a speaker. In other words it may be necessary for the system to 'over-act.' Anecdotally, the example of R2D2 (Lucas, Kurtz, Williams et al. op. cit.) comes to mind as an example where extreme prosodic manipulation seems to communicate liveliness quite effectively (see CD Track 14). This view is supported in a related context by Eide, Bakis, Hamza et al. who recount the need to collect exaggerated performances of expressive human speech in order to reliably reproduce the expression subsequent to the synthesis process (Eide, Bakis, Hamza et al. op.cit. p.221). The notion of setting more extreme values to the eventual set of paralinguistic/prosodic modifiers is addressed in subsequent tests. It may also indicate that the linguistic content is critical to the user perception of liveliness. This potential outcome is considered in 6.2.6 and in the framework implementations in chapter 7.

A weakness of the test is the comparative aspect of the evaluation. In other words, while it may be confidently expected that a user can correctly distinguish the three speech styles when comparing one speech style with another, this may be more difficult when speech-style is presented in isolation, without a reference point. To address this, a comprehensive range of testing methods, designed to triangulate results, are documented in ensuing chapters in this research. These include further conventionally-structured user evaluations using questionnaires and interviews in a comparative context, as well as evaluations designed to elicit responses at a subliminal and less-conscious level without a strict comparative context.

Another weakness was that neither of the studies directed the researcher to a single term that would reliably capture the improvement to synthetic speech proposed in the research objectives. The solution adopted was to present the user with a number of descriptors drawn from both studies which in combination would capture the notion of 'liveness'. Therefore, ensuing studies are targeted at identifying manifestations of the composite term 'liveness'; however, various descriptors are used to present this concept clearly to the participants. They are:

1. Live (see 0, 6.1.7)
2. Sounding spontaneous (see 6.1.8, 6.1.10, 6.1.9)
3. The difference between reading and speaking (see 6.1.8, 6.1.10, 6.1.9)
4. The best actor (see 6.1.8, 6.1.10, 6.1.9)

Before reviewing the studies in liveness, the next chapter will examine the sources and perspectives for the heuristics and metrics applied to the automation of the low level paralinguistic/prosodic modifiers designed to engender liveness in the PAT software tool.

5 Automated paralinguistic/prosodic modification

'...while the creation of the "best" algorithms for demonstrating human capabilities is a fascinating area of research, designers must always consider whether the "boring" approach of simply having a person do it by hand or faking an actual ability might yield better outcomes for the interface and the user.' (Nass & Brave *op. cit.* p.153)

In order to test the validity of the synthetic speech modifiers identified in the review of sources and perspectives in chapter 2 and to meet the requirements of proposition one, each feature has to be rendered automatically in an appropriate synthesis system. This must be done both in isolation, and in combination with other features. A methodology for defining the heuristics and metrics for each feature has to be derived and a system, with a generalisable synthetic voice implementation, built to host the tests. The PAT software tool facilitates automated paralinguistic/prosodic manipulations to a SSML (Synthetic Speech Mark-up Language) compatible speech synthesiser. The input data is 'plain text' and the resulting limitations on the possible automated prosodic/paralinguistic modifications are set out in this chapter. The sources and perspectives determining the heuristics and metrics applied to the paralinguistic/prosodic modifiers implemented in the PAT software tool, briefly reviewed in chapter 2, are explained in greater detail.

Although this chapter is a further elaboration on some of the literature identified in Chapter 2, it differs in that the sources identified for more detailed analysis all give rise to values and metrics which can be adapted for automated implementation in subsequent studies.

The studies are of two types:

1. Studies in which the user could iteratively manipulate the values assigned to the paralinguistic/prosodic modifiers until the speech output match their requirements.

2. Studies in which pre-assigned values were passed to the speech output for subsequent user evaluation.

The PAT software tool is designed to facilitate experimentation with the speech audio stream from a SSML (Synthetic Speech Mark-up language) compatible TTS system.

The PAT software tool controls the speech output and operates on the ‘time’ dimension, manipulating the auditory properties of the speech-stream output by the synthesiser as a serial stream of sequential SSML mark-up instructions. The standard speech (see structured vocabulary) can be provided by any SSML compliant speech synthesis system. The modifications rendered are labeled by the researcher as ‘synthetic voice acting’. The ‘synthetic voice acting’ is the product of the selected temporal variations (pauses and periodic rate changes), further modified by random factors and additional additive features (breaths and background sounds). These are identified in section 2.7. The metrics for items 1 – 5 (in List 4) are largely derived from linguistics. The metrics for item 5 are derived from Rabin’s theory of filtered randomness (Rabin 2004) and humanisation algorithms. The implementation of Item 6, background sounds, is not supported by sources or perspectives with quantifiable features; consequently, analysis of this feature is held over until Chapter 7.

1. Grammatical pauses
2. Non-grammatical pauses
3. Periodic tempo variations
4. Breath sounds
5. Random perturbations to the speech stream
6. Background sounds

List 4: Paralinguistic/prosodic modifiers implemented in the PAT software tool

The PAT software tool interface provides the researcher with means to construct automated multi-layered SSML edits, using the paralinguistic/prosodic modifiers listed above, to any plain text input or file. The edit can then be spoken by the synthesis system, allowing the researcher or focus group to make further modifications before deployment in an experiment or in the field; for example, in a web based survey or an arts installation. All the experiments and studies were dependent on detailed analysis and subsequent encoding of audible modifications to the speech-stream. This was facilitated by the PAT software tool.

Coinciding with the development, testing and evaluation of the PAT software tool, was the evolution of the PAT framework, including the representation of the 'Place' and 'Authenticity' dimensions and the setting and scripting rendering methods. The PAT tool was central to all the PAT framework tests in providing the source for the modified speech stream, either in near-real-time or as a recorded sample. In the following three chapters, for the sake of clarity, an analysis of the development of the two stages - the PAT software tool and the PAT framework - is presented sequentially. In fact, the two stages developed simultaneously, with significant cross-fertilisation.

Before commencing a detailed analysis of the process of acquiring and implementing metrics for the PAT tool it is useful to examine in more detail the linguistic concepts which underpin items 1 – 3 of List 4. These concepts fall broadly into an area of research defined in linguistics as prosody.

5.1 Paralinguistic/prosodic modifiers: perspectives from linguistics

'Prosody', according to Wennerstrom, "is a general term encompassing intonation, rhythm, tempo, loudness and pauses, as these interact with syntax, lexical meaning, and segmental phonology in spoken texts" (Wennerstrom op. cit. p.4).

Research into the prosody of English is extensive, and shares common ground with other branches of linguistics such as phonetics (the study of speech sounds), phonology (the study of relations between speech sounds in a particular language or in all languages) and pragmatics (the study of speech usage in everyday life). There is increasing awareness of the relevance of prosody analysis to discourse analysis; however, there are orthographic limitations to the accurate notation of prosody⁴⁰.

The sections on linguistic research on pauses in spontaneous speech draw heavily or cite directly from a series of studies by (Goldman Eisler op. cit.). The metrics cited, refer to specific

⁴⁰ Difficulties in the notation of prosody also present problems for current TTS systems which rely entirely on orthographic information in order to render the appropriate prosody.

experiments documented by Goldman Eisler. Space does not permit the full contextual analysis provided in the source and the interested reader is referred to the original text.

5.1.1 Pauses

Contrary to what Wennerstrom says, a pause is not, strictly speaking, a prosodic effect as it is segmental, applied serially, rather than applied simultaneously with the segmental phonology as other prosodic effects do. It may also be regarded as a paralinguistic phenomenon, as it operates outside the communicative optimisation remit applied by the speaker to other features of prosody. Pauses in speech have conventionally been categorised as either ‘filled’ or ‘unfilled’ (voiced or unvoiced). They may also be filled by a breath. Voiced pauses may take many forms, but the most frequently occurring type is an unrounded central vowel, such as “er” or “um”. Drawled speech characterised by lengthened phonemes could also be a form of pausing. Functionally, there is an overlap between the silent and voiced pause in terms of both distribution and duration. As a clarification of the function of a voiced pause, Crystal says ‘...some such definition of voiced pause as “capable of being substituted for by silence” would be necessary to avoid bringing too much vocal effect into the prosodic fold’ (Crystal op. cit. p.167). Thus in this research a silence may be regarded as broadly equivalent in function to a voiced pause. The synthesis of voiced pauses in a TTS system is outside the scope of this research.

When reading texts, depending only on familiarity with the material, the speaker’s pauses will be almost entirely aligned to the grammatical structure. The speaker will not only acknowledge the punctuation as a delimiter, but will process the other grammatical junctures listed below some not accessible to current synthesis models⁴¹.

The following pauses are designated ‘grammatical’ by Goldman Eisler⁴²:

1. Natural punctuation points.
2. Immediately preceding a conjunction (“and”, “but”, “therefore”, etc.)
3. Before relative and interrogative pronouns - (Relative (“that”): *This is the house // that Jack built.*) (Interrogative (“what”): *I asked you // what is going on?*)

⁴¹ The inaccessibility is due to limitations in current automated prosodic modelling reviewed in (5.3).

⁴² Examples added by the researcher and gathered from Wikipedia and other anonymous web sources.

4. When a question is direct or implied.
5. Before all adverbial clauses of time, manner and place - (*Time: Her father died // when she was young.*) (*Manner: I was never allowed to do things // the way I wanted to do them.*) (*Place: He said he was happy // where he was.*)
6. When complete parenthetical references are made.

There is evidence in spontaneous speech that pauses can take on a structural role, related to grammatical contexts and an unstructured role, such as hesitations that are unrelated to grammar and unpredictable.

The following pauses are designated 'non-grammatical' by Goldman Eisler⁴³:

1. Where a gap occurs in the middle or at the end of a phrase – (*In each of // the cells of the body //...*).
2. Where a gap occurs between words and phrases repeated – (*The question of the // of the economy*).
3. Where a gap occurs in the middle of a verbal compound – (*We have // taken issue with them and they are // resolved to oppose us*).
4. Where the structure of a sentence was disrupted by reconsideration or a false start – (*I think the problem with de Gaulle is the // What we have to remember about France is...*)⁴⁴.

According to Goldman Eisler's studies in spontaneous speech, 55% of all pauses occurred at grammatical junctures and 45% in non-grammatical places. Based on this information, we cannot conclude that unstructured, non-grammatical pauses are redundant. Non-grammatical pauses may be used paralinguistically to colour the speech, and thus to facilitate better communication. Nor can we assume that all the available grammatical pauses will be presented; only that in spontaneous speech almost as many pauses are likely to occur outside the grammatical framework as inside.

We may note that grammatical and non-grammatical pauses are commonplace in spontaneous speech. Although, as previously indicated, the term 'spontaneous' is not applicable in the context of user evaluation of synthetic speech, it is an appropriate and useful analogy of

⁴³ Examples from Goldman Eisler

⁴⁴ It may be assumed that any other occurrence of a pause that lies outside either of the lists may also be categorised as non-grammatical.

'liveness'. Accordingly, the patterns of pauses found in the linguistic literature detailing spontaneous speech are relevant to the encoding of liveness in synthetic speech and may provide useful guidelines for the development of appropriate heuristics and metrics. Whether we regard this as an additive or subtractive process - are we *adding* 'white space' or *removing* congestion in the speech stream - does not change the comparative simplicity envisaged in implementing additional empty pauses in synthetic speech. Audible breaths on the other hand are clearly additive. An examination of a comparatively simple paralinguistic effect (item 4 in List 4) is the subject of the next section.

5.1.2 Breaths and filled pauses in spontaneous speech

An actor is taught to breathe in order to conserve her voice and to project the sound in a large performing environment. Other than in specific forms of performance (see 2.5.6), particularly singing (see 2.5.5), performers are not instructed when they should breathe; this is left to their discretion or instinct⁴⁵. However, in normal speech, it appears that the temporal location and number of breath pauses is a function of the type of speech whether spontaneous or read.

As detailed in 2.4.2, in spontaneous speech there are many more pauses than in readings. A much higher proportion of the pauses taken in spontaneous speech are non-breath pauses. This is shown in Table 13 (Goldman Eisler op.cit.).

Pause type	Readings	Spontaneous speech
Non-breath pauses	30 (22.6%)	261 (65.9%)
Breath pauses	103 (77.4%)	135 (34.1%)
Total	133	396

Table 13: Breath pauses and non-breath pauses in spontaneous speech and readings. From a study by Goldman Eisler.

Pauses for breath tend to occur at grammatical points when the speaker is most in control, such as when reading a pre-prepared text, and at ungrammatical points in spontaneous speech. This is shown in Table 14. (ibid.)

Pause type	Readings	Spontaneous speech
------------	----------	--------------------

⁴⁵ The teaching of elocution (out of fashion today) included instruction on breathing.

Breaths at grammatical junctures	103 (100%)	93 (68.5%)
Breaths at non-grammatical junctures	0 (0%)	42 (31.5%)
Total	103	135

Table 14: Breath pauses at grammatical and non-grammatical junctures. From a study by Goldman Eisler.

There are a number of different types of filled and breath pauses in speech; some would require complex synthesis:

1. Inhalation, in which the voice output ceases and inhalation is heard.
2. Exhalation, in which the voice output ceases and exhalation is heard.
3. Nasal inhalation/exhalation; this is not heard often but has a highly intimate effect.
4. Involuntary vocal noises, like belches, coughs, etc.
5. Vocalised sound effects.
6. Other non-specific physiological vocalisations from the tongue, lip, teeth, or saliva.
7. Pauses filled by other external sounds.

Breath pauses are additive in so far as they require encoding into the speech stream. In the context of this research, this is a less desirable approach than subtractive or distractive processes, and, accordingly their introduction should be subject to caution (see 2.2 and 7.1.2). However, the synthesis of breaths has been shown to make a small improvement to recall (Whalen, Hoequist & Sheffert op. cit.) even when the breath sounds are not realistically aligned with the synthetic voice, and thus breaths may offer potential as a paralinguistic/prosodic modifier designed to evoke liveness unconstrained by realism.

The relation between breath and non-breath pauses in fluent and hesitant speech (the subject of the next section and item 5 in List 4) provides an additional potential paralinguistic/prosodic modifier for inclusion in the PAT software tool.

5.1.3 Periodic tempo (speech rate) variations

In spontaneous speech, pauses or hesitations interrupt the steady prosodic flow in a very complex, dynamic and interactive way. Despite the complexity, the periods of hesitant speech seem to relate to the periods of fluent speech, producing a periodic pattern of alternating

fluency and hesitation. It seems that during hesitant speech the speaker is mentally preparing for fluent speech⁴⁶. This is illustrated in Figure 17 (Goldman Eisler op.cit.)

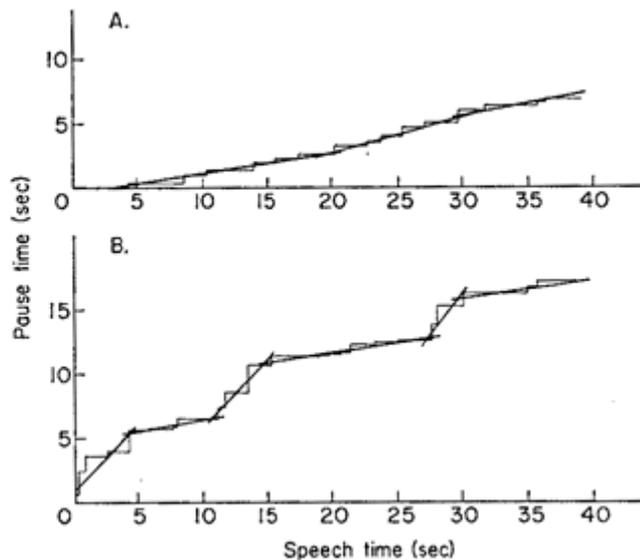


Figure 17: Temporal patterns in speech. A: reading; B: spontaneous speech.

Listening to spontaneous speech (CD Track 44) will readily reveal the pattern found in the lab by Goldman Eisler as a series of speech segments of indeterminate length and speech rate, alternating between fast and slow. The system is perceived as volatile and unpredictable; however, some metrics emerging from Goldman Eisler's studies into speech rates in spontaneous speech are listed below:

1. Length of utterance can determine speech rate. Long utterances produce a faster speech rate, while short utterances are treated more variably by speakers.
2. Approximately 200 syllables per minute is the most frequently occurring speech rate⁴⁷, and the fastest articulation rate varies between 6.7 and 8.2 syllables per second.
3. The articulation rate revealed in the studies occupied a range between 4.4 and 5.9 syllables per second.
4. The speech rate variation coefficient revealed in the studies was 21.5%.
5. The articulation rate variation coefficient revealed in the studies was 9.1%.

⁴⁶ Why this pattern occurs, is not universally agreed. Mental preparation is just one theoretical possibility.

⁴⁷ Speech rate according to Goldman Eisler "...is measured in the number of syllables a minute of the whole utterance. Articulation rate as the number of syllables a minute of the time spent in vocal activity." P. 24

The rate of speech production proved to be a function of the proportion of time taken up by hesitation pauses, and what is experienced as an increase in speech-speed proved to be a variation in the amount of pausing. It has also emerged that there is a relationship between pause time and the production of what Goldman Eisler refers to as 'rhythmic speech', which is speech where there is a significant relationship between the periods of hesitation and the periods of fluency. A less-than-30% proportion of pausing in a speech segment seems to preclude conditions favourable to the production of such temporal patterns.

From these studies it may be perceived that an intimate interrelationship between pauses and speech-rate exists. Pauses not only signal the distinction between spontaneous and read speech, but they are the prime factor in rendering the distinctive pattern of periodic speech-rate variation that is also a signifier of spontaneous speech.

5.1.4 A summary of pauses from the linguistic literature

The average distribution of pauses is a function of the grammatical structure of sentences (Goldman Eisler *ibid.*) but contextual issues related to the complexity of the material (be it prepared or unprepared), the experience of the speaker, and performative issues (such as 'dramatic' pauses) make the application of pauses in any utterance very individual. The range applicable to pause-type, distribution and duration that may occur is very large. This presents a highly variable picture presented in the following list based on the specifics of Goldman Eisler's studies⁴⁸:

1. Familiarity with verbal material results in fewer pauses.
2. Familiarity with verbal material results in shorter pauses.
3. Pause lengths are determined by the type of situation in which the speech is uttered.
4. The relative duration of pauses in speech is between 4% and 67% of overall speech time.
5. Pauses in discussions are never longer than 3 seconds and 99% are less than 2 seconds.
6. Breath pauses last between 500 milliseconds and 1 second.
7. Speakers take between 2⁴⁹ and 20 respirations per minute.
8. Breathing takes between 2.5% and 25% of speaking time.

⁴⁸ Speech produced in different contexts and with different speakers.

⁴⁹ Two breaths per minute seem remarkably few. The statistic is reproduced from Goldman-Eisler p.23.

9. Fluctuations in speech rate are a function of the duration and frequency of empty pauses to a much greater extent than they are of breathing pauses.

List 5: A summary of pauses in speech (from Goldman Eisler)

Some factors in List 5 present a suitable set of ranges upon which user permutations for the paralinguistic/prosodic modifiers may be constructed, but also demonstrate the need for some heuristics based on randomly generated values.

If 'spontaneity' is an appropriate analogue for 'liveness' then one approach would be to apply the pattern of pauses found in spontaneous human speech to the standard speech produced by default by the synthesiser and test for increased 'liveness'. Before considering the implementation of such a proposal, a brief discussion of the difficulties presented in computing prosody for synthetic speech is the subject of the next section.

5.1.5 Prosody in synthetic speech is a complexity problem: a discussion

The view that language is a complex set of interrelated systems that cannot easily be separated out has parallels with claims that the Turing Test is a complexity problem. Human intelligence is a set of interrelated dynamic systems, of which language (one of a number of features of human intelligence tested by the Turing test) is one manifestation. The difficulty in passing the Turing test is not the act of rendering language, per se; it is the act of supporting basic language capabilities with the other systems of knowledge, which are easily accessible to humans but largely inaccessible to machines. This is known by the AI community as 'World Knowledge' (Dennett 2004). Speech is a manifestation of language, itself a manifestation of human intelligence and prosody a manifestation of speech. To expect to synthesise prosody without first synthesising human intelligence may be unrealistic because it is not possible to isolate the constituent variables that give rise to credible prosody (and these are not universally agreed) from 'World Knowledge'. Hence, at present, the accurate automated production of credible prosody in synthetic speech is proving to be an intractable problem⁵⁰. Paralinguistic/prosodic modifiers exist in commercial speech systems, but they are designed for

⁵⁰ This view was confirmed by a conversation with Alan Black, the co-inventor of the Festival Speech System, now of the Language Technologies Institute at Carnegie Mellon University. He agreed that the Turing Test was unlikely to be passed at least in his lifetime, and that the best prosody likely to be attained in the foreseeable future will not sound realistic (Conversation with Alan Black 13/03/09, University of York, Computer Science Dept).

the generation of prosody in predetermined utterances. These require significant human intervention, and are not particularly convincing in a long speech segment when knowledge at greater-than-sentence-level is required by the synthesis model. Thus, a comprehensive prosodic approach to generating liveness in synthetic speech is a complexity problem akin to other problems addressed by complexity theories.

Complexity theories have been successful in gaining deeper understandings of natural systems. In particular, they have been successfully applied to non-linear natural and human systems in which small changes to input may lead to a large unpredictable change in output; systems such as the weather, population growth and the stock-market. The contribution from complexity theory is to examine subhuman processes at superhuman scales and, in so doing, seek to reveal simpler generic process models. At the subhuman-level human prosody is inflected by a multiplicity of complex interrelated variables that are noncomputable (see footnote 20) at present. Rather than use the term subhuman Franchi and Güzeldere use the term microworld but make the same point from an AI perspective:

“At present, it is essentially a received view that trying to circumvent the ungeneralisability problem by adding more microworld constructions results in nothing but ill-tamed complexity of a sort that ultimately runs up against an insurmountable wall (Franchi & Güzeldere *op. cit.*).

At a superhuman-level, with much of the detail subsumed, the patterns of prosody may readily be perceived as cycles of hesitation and fluency (pausing and not pausing). By computing the superhuman-level features into a synthetic voice and evoking liveness and WSOD, the significance for the user of the difficult subhuman levels may be diminished. The viability of this proposal is explored in subsequent studies.

A deeper examination of complexity theory (Lewin 1992) and its relationship to natural systems that could include speech is a fascinating prospect but is outside the scope of this thesis.

5.2 Paralinguistic/prosodic modifiers

In 2.4.1, paralinguistic modifiers were identified as an alternative source of temporal variation that may give rise to 'liveness'. The distinction between 'paralinguistic' and 'prosodic' is not clear, and, for the purpose of this research, the distinction is determined according to the predicted effect on the user's semantic perception of the speech. It is important that the modifiers do not have an adverse effect on the intelligibility of the synthetic speech; nor should they disturb the user's semantic interpretation of the speech. This dilemma is appropriately summed up by Monaghan: "If in doubt, make sure the prosody is neutral and let the user decide on the interpretation"(Monaghan 2002). Although the notion of 'neutral' speech (see structured vocabulary) is problematic, Monaghan rightly advises caution, and it is intended that the modifications outlined herein elicit the user's own semantic interpretation rather than impose one. This is a rather critical point, and needs emphasis. Current research is directed to the twin goals of intelligibility and appropriate expressiveness, within a rigid framework of verisimilitude to human speech. The intention is that as far as possible the meaning encoded as the input is analogous with the meaning decoded as the output. The paralinguistic modifiers presented in this research are bound to encode meaning less reliably than prosodic modifiers. Thus, if the objective was comparable with that stated above (intelligibility and appropriate expressiveness), paralinguistic modifiers would be sure to fail. Happily, that is not the objective.

Of the range of modifiers, speech rate, and pauses offer the potential to be the least semantically disturbing. In other words, users accept reasonable variation in rates of both speech and pausing among different speakers, and do not attach undue semantic significance to the variation. Goldman Eisler provides evidence that an utterance spoken quickly and the same utterance spoken more slowly do not have appreciably different meanings to the listener. Goldman Eisler concludes that the rate of articulation, though highly variable across different individuals, is a personality constant even when the cognitive tasks are very different (Goldman Eisler op .cit. p.25). From personal experience, we know that the effect of a person articulating rapidly will not substantially change the meaning communicated. Similarly, an utterance interrupted by a short hesitation will not be interpreted significantly differently to an uninterrupted utterance. This claim has its detractors, such as Fónagy who says "The common

link between speech and silence is that the same interpretative processes apply to someone remaining meaningfully silent in discourse as to their speaking” (Fónagy 2001, p.598 citing Jaworski). However, exercising interpretative processes that may evoke WSOD is in line with our objectives. We only need avoid interpretative processes which arrive at a misleading conclusion, thus undermining intelligibility or causing semantic disturbance. We are reminded that hesitations are a reliable indicator of spontaneity in speech, but the literature provides no evidence either way for the impact on semantics. Therefore, it cannot be confidently claimed that pauses create no semantic disturbance; only that the other prosodic variables, such as a change of intonation, loudness or rhythm, could change the intended meaning of the utterance significantly more.

According to sources and perspectives outlined in Chapter 2, based on a plain text input, it may be possible to generate mark-up for speech-rate settings and pauses automatically and/or randomly to simulate liveness. However, one of the most significant limitations on the development of prosodically enhanced TTS is the dearth of prosodic information encoded in plain text. This is the subject of the next section.

5.3 Prosodic representation and mark-up

The technological platform utilised in this research transforms data represented as plain text into synthetic speech. This is the basis of a text-to-speech-system. TTS prosodic representation is captured in the same text-representation as the literal text-content (it is represented as ‘to be spoken’). Prior to any automated mark-up, the text will be prepared using a word processor or similar system. The text is not assumed to be encoded with any additional prosodic signifiers. The only pauses readily encoded in the text are grammatical. Section 5.1.1 documented Goldman Eisler’s designation of two types of pauses: ‘grammatical’ and ‘non-grammatical’. Her designation of ‘grammatical’ pauses is reproduced here for convenience:

1. Natural punctuation points.
2. Immediately preceding a conjunction (“and”, “but”, “therefore”, etc.)
3. Before relative and interrogative pronouns - (Relative (“that”): *This is the house // that Jack built.*) (Interrogative (“what”): *I asked you // what is going on?*)

4. When a question is direct or implied.
5. Before all adverbial clauses of time, manner and place - (Time: *Her father died // when she was young.*) (Manner: *I was never allowed to do things // the way I wanted to do them.*) (Place: *He said he was happy // where he was.*)
6. When complete parenthetical references are made.

List 6: Grammatical pauses according to Goldman Eisler

In List 6 we may assume that the synthesiser will only render orthographic pauses represented by item (1), 'natural punctuation points'⁵¹, and item (4), 'direct' questions (when indicated with a question mark). To render the other grammatical pauses would require the inclusion of a model in the synthesiser for 'parts of speech' tagging as proposed by (Veronis, Christo, Coutois et al. 1997). The researcher knows of no TTS system capable of generating credible prosody based on the full syntactic parsing of an arbitrary plain text input. Certainly this has yet to be included in the Microsoft SAPI implementation used in the studies.

Only information that can be represented as plain-text, or marked up automatically, is of relevance to this research; therefore, much that is afforded attention by linguists cannot be included. In particular, information encoded in speech segmentally (the smallest practical unit of speech sound) or suprasegmentally (information like tone and stress) cannot be orthographically represented to the TTS system in plain text. While it is possible to send instructions to the synthesiser to modify the output phonetically - for example to correct mispronunciations - this requires human intervention, including the provision of mark-up, and does not address the main objective expressed in the first proposition set out in 1.1.10 for the 'automated modification of pauses and speech rate variations'.

To provide clarity for the reader, the linguistic features that can be represented in plain text require a set of precise lexical references, as well as a simple label that can be interchanged between speech data and text data. This is made difficult by some variation in the usage of key concepts and terms distributed through the literature.

The linguistic components that can be represented as plain text are presented in Table 15. The labels applied are original to this research.

⁵¹ This is subsequently tested in establishing the Microsoft SAPI benchmark settings in Section 6.1.3

Label in this research	Linguistic reference	Definition	Example
Word	Word	“A unit that is typically larger than a morpheme but smaller than a phrase” (Trask op. cit.).	“Synthesiser”
Space	Double word boundaries	Boundaries between two consecutive words.	“The synthesiser”
Punctuation	Orthographic features	Graphemes such as capitalisations and punctuation marks.	The synthesiser was astounding!”
Utterance	Utterance	Any single piece of speech. Normally a sentence.	The synthesiser was of poor quality.
Paragraph	No linguistic reference	A large unit of speech as defined by a paragraph of text.	Editorial options regarding the duration of pauses were the same as grammatical pauses with a choice between random duration within user defined limits or user defined constants.

Table 15: Linguistic components that can be represented as plain text

The next step is to consider the prosodic and paralinguistic variables that could be marked up with minimal semantic inference. Thus, the full list of prosodic and paralinguistic variables considered in this research are shown in Table 16.

Label in this research	Linguistic reference	Definition	ASCII Code	SSML Code (as implemented in the Microsoft SDK)
Grammatical pause	Orthographic features	Punctuated boundaries between two consecutive words. In these studies this was limited to full stops, question marks and commas.	ASCII period (.) ASCII comma (,) ASCII question mark (?)	<silence msec="string" />
Non grammatical pause	Double word boundaries	Boundaries between two consecutive words with no punctuation.	ASCII space ()	<silence msec="string" />
Paragraph pause	No linguistic reference	A segment of speech contained in one orthographically-defined paragraph.	ASCII tab () ASCII form feed ()	<silence msec="string" />
Breath	Filled pause (with breath)	A synthesised breath augmenting or substituting for a grammatical or non-grammatical pause. Non-concurrent with the speech stream	NA	The following SSML mark-up can be used to play audio files including breaths or background sounds ⁵²
Sound	Sound from elsewhere	A sound additional to the speech stream that may be heard concurrently with the speech stream; or augments or substitutes for a grammatical or non-grammatical pause.	NA	<ssml:audio src="string" />.
Speech-rate	Articulation rate	Variation in the rate of speech production	NA	<prosody rate="+/- (n)%"></prosody> ⁵³

Table 16: The list of prosodic and paralinguistic variables considered in this research

⁵² In the PAT software tool this method was found to have efficiency constraints consequently, the following SSML mark-up was used: <bookmark mark="string"/>. The bookmark value was passed to the synthesiser triggering the appropriate audio file (a breath or a background sound) to play. To play the sound concurrently or non-concurrently further synthesiser specific instructions must be passed.

⁵³ In the PAT software tool speech rate changes were executed as low level instructions to the synthesiser although the SSML method should work just as well

Based on the preceding analysis the scope of the first proposition of this thesis may be reexamined and represented as follows:

1. Prosody in human speech is not yet fully understood. No comprehensive representation models exist.
 - 1.1. Therefore, credible TTS prosody based on human models is unlikely, at present.
2. The TTS system must process plain text.
 - 2.1. Very little prosodic information can be encoded as plain text (see section 5.3).
3. The TTS system must process any text in the host language with no appreciable delay.
 - 3.1. Suitable controls do not exist to automate prosody in a real-time synthesis system.
4. Some prosodic modulations can modulate meaning.
 - 4.1. Therefore, the modulations must be selected that have least impact on the perceived meaning of the words.
5. Pauses and speech-rate variations are prosodic features that have less impact on semantics than intonation, rhythm and loudness and map easily to human paralinguistic/prosodic modifiers found in the encoding of spontaneous speech when compared to pre-prepared speech.
 - 5.1. Some features of spontaneity may be said to be analogous with 'liveness'.
6. Thus pause and speech rate variation are the most viable candidate to encode 'liveness' within the constraints of current TTS technologies.

5.3.1 Automated mark-up in the PAT software tool

The PAT software tool works with the synthesiser's built-in parsing of plain text, in addition to the SSML mark-up generated automatically. So that the tests were fair, both the synthesiser's characteristic parsing of plain text and SSML mark-up were transparent to the researcher and documented in this thesis⁵⁴. Using a number of different synthesisers would result in an unmanageable increase in the number of independent variables requiring testing, and the likelihood that participants would evaluate the fundamental quality of the voice synthesis

⁵⁴ Despite the SSML standard, individual systems will implement SSML mark-up differently. Provided the system makes appropriate provision to ignore mark-up it cannot process, SSML compliance can still be claimed.

rather than the subtle modifications to the paralinguistic/prosodic variables set out in this research.

The Microsoft synthesiser was chosen as the platform with the most robust SSML compliance. The decision was made to build upon the Microsoft synthesis model rather than modify it. For example, the Microsoft synthesiser is silent for 300 milliseconds (ms) at a full stop and 250 ms at a comma. The PAT software tool appends those values with additional values or leaves them as they are, rather than replacing them. In theory, the tool could be used to supplant those values but the result would be to progressively chip away at the Microsoft synthesis model until a different model emerges. This could confuse the results. The goal implied in this research and set out in the propositions is not to build new synthesis models but to improve existing ones, and that means augmenting existing default settings rather than supplanting them.

The default settings for the Microsoft synthesiser used in the studies are not published and accordingly an acoustic analysis of the default settings for the synthesiser used in this research is presented in 6.1.3.

5.3.2 Randomness in automated mark-up

In the PAT tool, randomness is used to simulate the unpredictable choices a human speaker may make about the distribution, duration and type of pause. In common with Meyer-Eppler's aleatoric concept (Meyer-Eppler 1957) this is set against a determinist framework. The determinist framework is based both on the range of variables set out in 6.2 and on choices made by users. Thus, participants can select from a determinist set of heuristically-determined variables, the final value of which may optionally be determined randomly. The objective in using aleatoric principles is to achieve randomness that, when experienced by the user 'feels right' and evokes 'liveness'. This principle is aligned to the notion of humanisation algorithms discussed in 2.5.8. A lower-level analysis of the notion of 'randomness that is perceived as appropriate', known as 'filtered randomness', is explored in the next section.

5.3.3 Filtered randomness in automated mark-up

The following section directly cites or is adapted from (Rabin 2004).

Filtered randomness evolved from studies that indicated that “...there was a big difference between what randomness can create and what humans will perceive as being random” (Rabin 2004, p.71). If the random disturbances to the prosodic flow do not appear to be random, but appear to be machine generated, the user will mistrust the genuineness of the output.

Rabin generates the numbers in Figure 18 using the C++ function `rand() % 2`. When focusing on a small section of numbers – for example, the sequence of eight “1”s (highlighted in Figure 18) - the user may start to question the randomness of the outcome.

```
010110001101000111101111000111011011011111110001010000100100001011010
0101000100100111101100001011010001000110000000010111101111101011011
0011011111100000011000001111111110011110111001101101101111011111010011
011011011110101010011000110111001010111000001000110001000100010100000
111101100000110100110010
```

Figure 18: 300 random bits generated by the C++ function `rand() % 2`

Rabin’s algorithms for filtering Boolean chance modify the first set of numbers to produce a second set by applying the following three rules based on psychological studies (Falk & Konold 1997):

1. Raise the number of alterations above the natural value.
2. Restrict unreasonable runs of the same value.
3. Eliminate recognisable patterns.

The results are shown in Figure 19.

```
010101000101000110110010100100101001011100101110101101001000110101011
101011101001100010101101100101010001100100101011101010111001000110110
101011000101100011010101101000110110110010100100011011010101100111001
101101010110111001110101101010011000101001101101010110001101110010101
001011100101110100101101
```

Figure 19: 300 filtered random bits

The idea of alterations is important and Falk & Konold found that score sequences were described as more random when a random head/tail sequence alternated between heads and tails at a higher rate than a true random result would have given. This is shown in Figure 20:

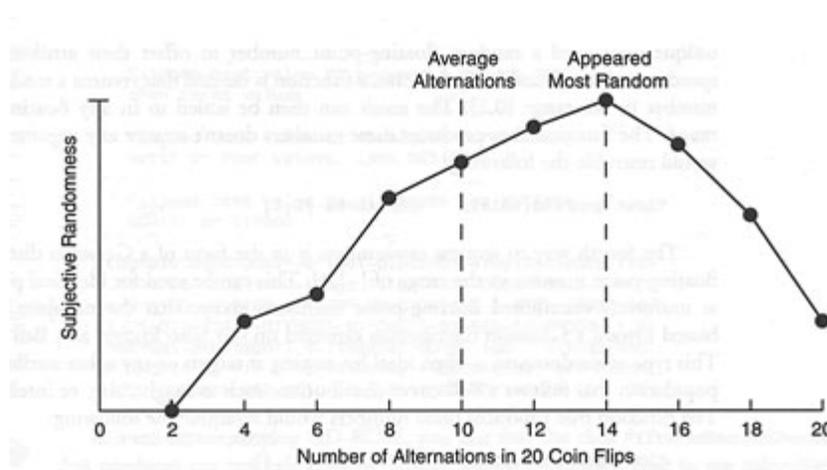


Figure 20: Subjective randomness

Given sequences containing 20 coin-flips each, test subjects rated the randomness of various sequences. Sequences that alternated 14 times between heads and tails appeared the most random

The PAT software tool offers a range of functions that trigger different types of paralinguistic/prosodic variation with different parameters. The output should appear random. This could be specified randomly, but Rabin suggests that this may result in user incredulity. Adjustments to the parameters based on a range of filtered random integers, generated such that the temporal occurrence and parameter settings are appropriately filtered, produce more believable results. The result is less random, but is perceived as more so.

Rabin's rules for the generation of random integers are as follows:

1. Restrict repeating numbers (this can be relaxed to allow limited runs).
2. Restrict the number of times a given value can appear in the last 10 values.
3. Restrict the number of values in a row that form an increasing or decreasing counting sequence.
4. Restrict the number of values in a row that are at the bottom of the range or at the top of the range.
5. Restrict two pairs of numbers in a row.
6. Restrict a motif of two numbers repeating immediately.
7. Restrict a motif of three numbers appearing in the last 10 values.

8. Restrict a motif of three numbers and its mirror from appearing in the last 10 values.

In the sequence shown in Figure 21, each time a rule throws out a number, the number is surrounded by a symbol indicating which rule is violated in the key. The original set of numbers was generated with `rand()%10`.

```
098(8)637410464-6-798526250758090-9-51[0]9743(3)17501
82095253608(8)67(7)18594281806586$5$025(5)059(9)27178
491972073780[7]5398137(7)61732(2)734186572856149(9)65
9290252809415808*2*-0-386[8]10595(5)-9-2(2)6085325350
[5]738186(6)2618(8)2[8]383(3)461(1)74312(2)8612976810
20$1$6(6)2595(5)60859808637274(4)7(7)1(1)[7]698925478
{9}381(1)350282-8-16908197581076{5}7-6-9[7]*1*(9)$8$0
130(0)$2$582(2)1673721*6*(1)95(5)365638703547(7)0(0)4
10(0)61039251(1)2(2)7436125710408(8)07398
```

Figure 21: Filtered random integers according to Rabin's rules

Key:

Rule 1: ()	Rule 2: []	Rule 3: { }	Rule 4: \$ \$
Rule 5: # #	Rule 6: - -	Rule 7: * *	Rule 8: * *

Randomness is a powerful tool for the simulation of a human, rather than machine, source. Readers are reminded of the humanness-algorithms described in 2.5.8. This said, true randomness must be constrained so that users perceive it as random. The application of the rule set for filtered random integers provides a solution to the automated mark-up of prosodic or paralinguistic variables with random values.

5.4 Conclusions

In this chapter, the sources and perspectives for heuristics and metrics applied to the prosodic and paralinguistic modifiers to be implemented in the PAT software tool have been reviewed in detail.

Valuable insights have been gleaned from linguistics, particularly research into spontaneous and read speech. The potential for integration of additional filled and unfilled pauses and

periodic speech rate variations into synthetic speech has been examined. An understanding of the implicit limitations of current TTS systems which rely on orthographic data included in plain text format to define prosodic variables has been shown. The prosodic features that can be represented in plain text and passed to a TTS system have been provided with a set of precise lexical references, as well as a simple label (to be used in this research) that can be used with speech and text data. A refinement to previous discussions on randomness has been revealed in Rabin's notion of filtered randomness which prevents actual randomness from arousing suspicion of non-randomness in users.

Based on this, a methodology to define a set of heuristics and metrics that could be implemented in the PAT software tool is required. The objective is not to provide the system with set values that may synthesise liveness in synthetic speech, but rather to provide the researcher and user with an appropriate range of variables that can be modified or randomised. It is important to avoid providing options for prosodic and paralinguistic modification that encourage absurd outcomes, like a three-minute pause at a comma. It is also important to address the difficulties users have of detecting and labeling more subtle changes, identified in the study documented in 4.2. This is done by providing ranges that allow for some extreme permutations.

Developing and implementing an appropriate methodology is the subject of the next chapter.

6 Evaluating the values for the paralinguistic/prosodic modifiers

“When we wish that the person we are speaking to to pay attention to our speech, or we want him to be struck by our thoughts that his heart should feel what we feel, we should separate the many and different ideas we present by evident moments of repose.”Francois Riccoboni in ‘The Art of Theatre 1750’ (Benedetti p.77 op.cit.)

This chapter documents the process of evaluating in a series of user studies the settings for the paralinguistic/prosodic modifiers implemented in the PAT software tool. Five studies were undertaken: a focus group study, three web-based pilot studies and a paper-based user survey. Two of the web-based pilot studies failed to produce reliable results. The other studies broadly supported the evidence from the literature, but the results directed the researcher to consideration of a broader context for the successful rendering of ‘liveness’; a context subsequently developed into the PAT framework.

6.1.1 Application environment

The system chosen for PAT framework testing was the TTS system included with the Microsoft Speech Application Interface (SAPI). The voices were Microsoft Mary and Microsoft Sam. The application was written in Microsoft Visual Basic, and prosodic instructions were passed to the synthesiser using Synthetic Speech Mark-up Language (SSML). The Microsoft system produces tolerably pleasant Americanised female and male voices, with the customary robotic twang associated with the synthesis technology employed. The TTS system produces a variety of errors in parsing, pronunciation and prosody, depending on the text passed to it. For all studies, parsing errors (usually misidentifications of special characters) were manually corrected; however, no corrections were made to the pronunciation or prosody, despite some obvious peculiarities. Although the Microsoft TTS system processes the complete SSML file

before initiating the speech output, modifications passed to the system can be processed rapidly enough for the real-time illusion to be presented to the user.

The benchmark settings for all the prosodic and paralinguistic variables accessible in the system were pre-set at the Microsoft default (see 6.1.3). The full set of prosodic variables modifiable with SSML is specified by the W3C (W3C 2005).

In addition to near-real-time processing, the system could be used to streamline the creation of audio samples for comparative evaluation tests. The editing and deployment process can be described as follows:

STEP 1: Input plain text.

STEP 2: Make a selection from the paralinguistic/prosodic modifiers at the interface.

STEP 3: Iteratively review the speech audio output.

STEP 4: Record the samples if necessary.

STEP 5: Evaluate the speech audio output with users or experts.

STEP 6: Deploy the rendered audio in the field or in an experiment.

Interface expansion was facilitated using MIDI. This provides options for deployment in the field with non-traditional interfaces and the use of MIDI sensors and other input devices. For example in 'please wait with me' (see 7.1.4), a user detected using a MIDI sensor could interact with the system via a specially adapted telephone keypad and could generate a randomly chosen speech from a set with pre-determined prosodic edits, together with background sounds and an audible output, in something like real-time.

Files could be passed to the system either manually, by opening a chosen file and cutting and pasting, or automatically. In addition, access to a choice of installed voices and features specific to the tests outlined herein is provided. The interface for the PAT software tool, showing the key paralinguistic/prosodic modifiers, is shown in Figure 23.

6.1.2 Application functions in the PAT software tool

Prior to establishing an appropriate set of heuristics for the paralinguistic/prosodic modifiers the PAT software tool was built as a host application. This section describes the functions incorporated into the tool.

Overview: Processing is facilitated by passing SSML mark-up to the synthesiser; thus, most of the function outputs are SSML marked up text. Direct interrogation of low level functions of the synthesiser is not provided at interface level but implemented in the code when required. The control window displays the SSML mark-up, providing visual feedback to the researcher as a useful secondary verification method. All functions are optional; a value of zero or non-selection of a control bypasses the function. A schematic of the PAT software is provided in Figure 22 to accompany the screenshot of the interface. The key to the schematic is List 7.

- (1) Unicode text input to the system.
- (2) The string is converted to an array of characters and spaces. Calculations are made of the number of words, spaces and orthographic features.
- (3) The results of the calculations are returned to the interface, providing the ranges of modifications available to the user.
- (4) The user makes the modifications they choose at the interface.
- (5) Optional data is input by the speech recognition interface (application dependent).
- (6) Optional data is input by the MIDI interface (application dependent).
- (7) Values are generated according to the user or external interface requests.
- (8) Optional filtered random values are generated.
- (9) The set of values is interpolated into the array and the array is converted to a string of SSML mark-up.
- (10) The SSML marked-up text is passed to the synthesiser.
- (11) Synthetic speech output.
- (12) The SSML mark-up is passed back to the user interface.

List 7: Key to Figure 22

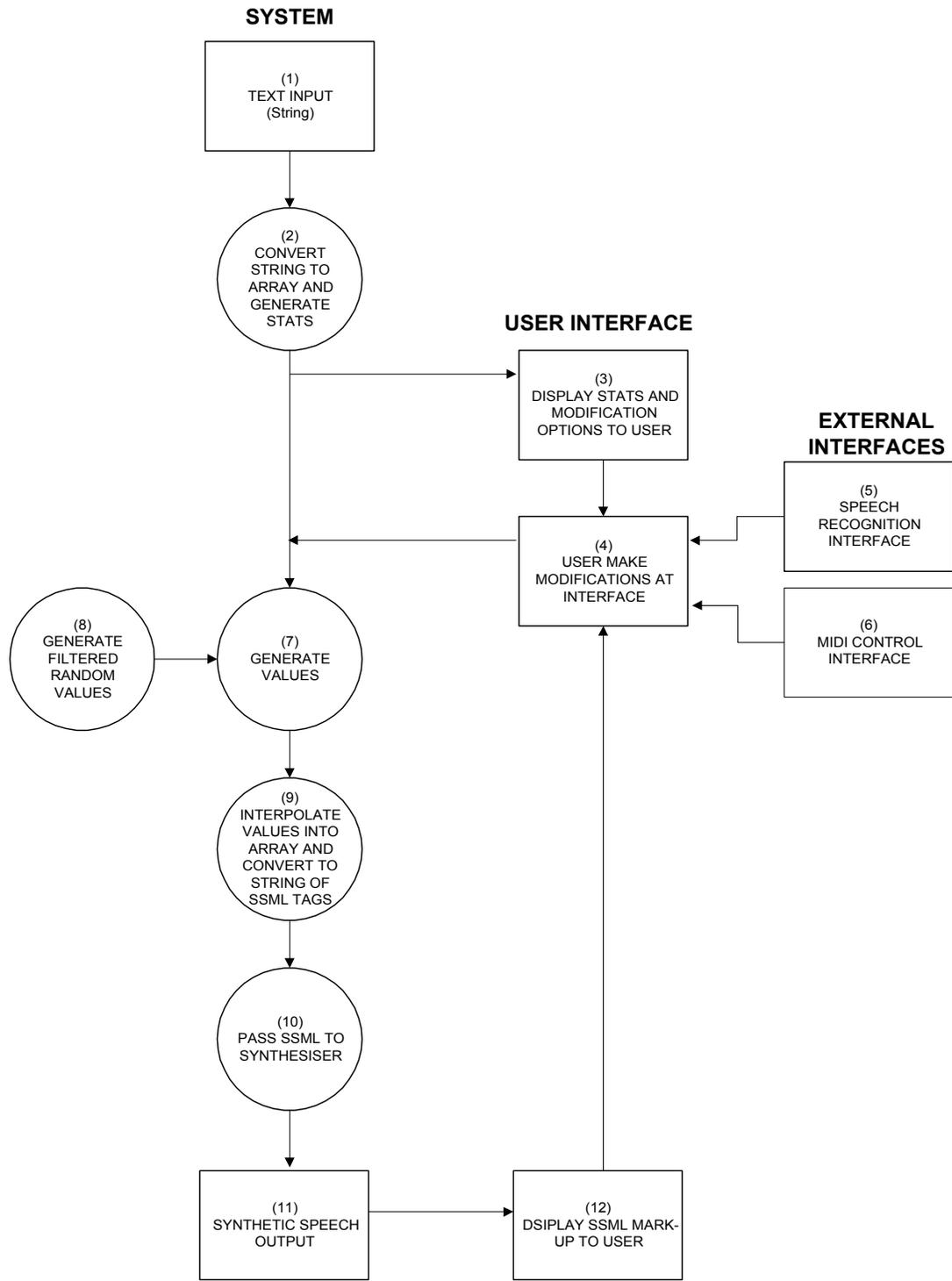


Figure 22: Schematic of the PAT software tool

Control functions: Files in Unicode '.txt' format may be opened either manually or by the system, depending on the implementation. No limit to file size was tested; however, subsequent operations on the file would be likely to fail with excessively long files. During the editing process, xml edits may be saved or discarded. Selecting the 'speak' control will output synthetic speech modified by any automated SSML mark-up specified at the interface.

Grammatical pauses: These are filled (with a breath sound) or unfilled pauses aligned to the orthographic markers (for example, '.', or ',', or '?'). The following rules were applied to the production of grammatical pauses:

1. Pause durations can be assigned either by setting a fixed value or selected randomly within a range of values set at the interface by the user.
2. The ratio of filled to unfilled pauses can be set at the interface by the user.
3. The distribution of filled to unfilled pauses within a chosen segment will be determined randomly by the system.
4. It is illegal for pauses to fall immediately before or after a pause (producing a cumulative longer pause). This includes before or after non-grammatical pauses.

In the PAT software tool, the procedure for processing grammatical pauses involved, first, the system importing a text, then editorial options set at the interface, regarding the orthographic feature indicating a pause (',' or ',') and the type (filled or unfilled) and duration of pauses, with a choice between random duration within user defined limits or user defined constants.

Non-grammatical pauses (periodic breath-like pauses): These are filled or unfilled pauses in the speech stream distributed according to rules that give the impression of random breath-hesitations, similar to those in spontaneous speech. The following rules were applied to the production of non-grammatical pauses:

1. Pause durations can be assigned either by setting a fixed value or selected randomly within a range of values set at the interface by the user.
2. The ration of filled to unfilled pauses can be set at the interface by the user.
3. The distribution of filled to unfilled pauses within a chosen segment will be determined randomly by the system.

4. The theoretical maximum number of non-grammatical pauses cannot be greater than the number of words less the number of grammatical punctuation points. (This theoretical limit is modified by the constraint imposed by item 7)
5. Pauses cannot fall immediately before or after a pause (producing a cumulative, longer pause). This included before or after grammatical pauses.
6. Pauses would be distributed as evenly as possible in the speech segment. The non-evenly distributed pause would fall at the end of the speech segment.
7. The number of non-grammatical pauses available for modification by the user is proportional to the number of words and cannot be greater than one pause every four words.

Item 6 is implemented to create the pulse-like effect described by Wennerstrom in Table 4 and to emulate a rhythmic predictability comparable to that found in both renaissance verse and in recitative (see sections 2.5.3 and 2.5.5).

In the PAT tool, the procedure for processing non-grammatical pauses was as follows. The system would import a text. The system would then calculate the number of individual words, less the number of grammatical punctuation points. This would produce the theoretical maximum number of non-grammatical pauses. The user would then set the number of non-grammatical pauses they wanted (with a maximum of one every four white spaces). The system would calculate an approximately-even-distribution for the non-grammatical pauses within the speech segment. Editorial options regarding the durations of pauses were the same as grammatical pauses, with a choice between random durations within user-defined limits, or user-defined constants. As some non-grammatical pauses would fall shortly after a grammatical pause, this sometimes produced a random effect, despite the even distribution.

Periodic tempo variations: Periodic tempo variations are applied at a user-specified marker. In all studies, this was a full stop. The implementation of tempo variation in the Microsoft synthesiser is not the same as that specified by Goldman Eisler in human speech (see section 5.1.3). Rather than modifying the duration of the pauses, the Microsoft system modifies both the pause duration and the articulation rate. Although this is an imperfect analogue of the human system, the effect is acceptable in the context of this research.

The periodic tempo variations were implemented in the PAT tool as follows:

The user could adjust the speed of alternate speech chunks of hesitant and fluent speech. No control was provided over which segment was hesitant and which fluent, only that the speed would be adjusted at the first designated marker and then readjusted at the subsequent marker, and so on, to the end of the text. The text, up to the first designated punctuation point, would be read at the default speed. The hesitancy rate could be adjusted from zero to (minus) - 20% and the fluency rate from zero to (plus) + 20% or both could be adjusted. The articulation rate variation from the Microsoft default was approximately 80% accurate, and should be taken as indicative. Subtle changes to the fluency and hesitancy setting produced a surprisingly natural effect, while extreme changes produced a grotesque, sea-sick effect.

Breath sounds: All pauses whether grammatical or non-grammatical could either be rendered silently or filled with a breath sound. The ratio of breath to non-breath pauses could be determined by the user, within a range of no pauses to all pauses filled with breath sounds. The sounds themselves were extracted from human recordings of readings and performances, and then processed to match the F0 of the synthetic voice. This was done by ear, using pitch modification tools in Steinberg Cubase™. Two sets of breath sounds were produced for the male and female voices. The system selected a breath sample consecutively from a looped list of 10 samples. There was no attempt to define breath types e.g. deep, long, snatched etc. but the effect was reported by users as convincing (an example is included on the CD Track 74). Latency could be adjusted by modifying the delay after the breath sound finished playing to create an overlap with the next utterance. All breaths had to have an associate pause rendered prior to them to give the best effect.

Actor pauses: Actor pauses are additional pauses indicated by a user-specified marker. In all studies, these were paragraph breaks. The following rules were applied to the production of actor pauses:

1. Values can be assigned either by setting a fixed value or selected randomly within a range of values set at the interface by the user.
2. The proportion of filled to unfilled pauses can be set at the interface by the user.
3. The distribution of filled to unfilled pauses within a chosen text will be determined randomly by the system.

Audio settings: These facilitate the setting of the audio output devices. For optimal performance, a multi-channel sound card is required, in order that the speech output and the non-speech (e.g. breath sounds) can be more carefully controlled.⁵⁵

Voice settings: Any installed SSML compliant voice can be manually selected or selected automatically by the system. The delay after a breath sound and before the voice resumes can also be set between 0 and 600 milliseconds.

MIDI settings: These facilitate the setting of MIDI input and output devices. This is required when extending the interface - to a telephone keypad, for example - as was done in the installations 'please wait with me' and 'Call Centre, both described in Chapter 7.

Other dialogues and controls: These provide additional interface options and information on the generation or archive files and user records. There would be no benefit to the reader in explaining these functions.

Call center, Call waiting, Not I screen and MMCP controls: These relate to specific implementations. There would be no benefit to the reader in explaining these functions.

⁵⁵ The sound devices used were M-Audio™ FireWire external sound cards '410' and 'Audiophile', with up to 8 separate output channels available.

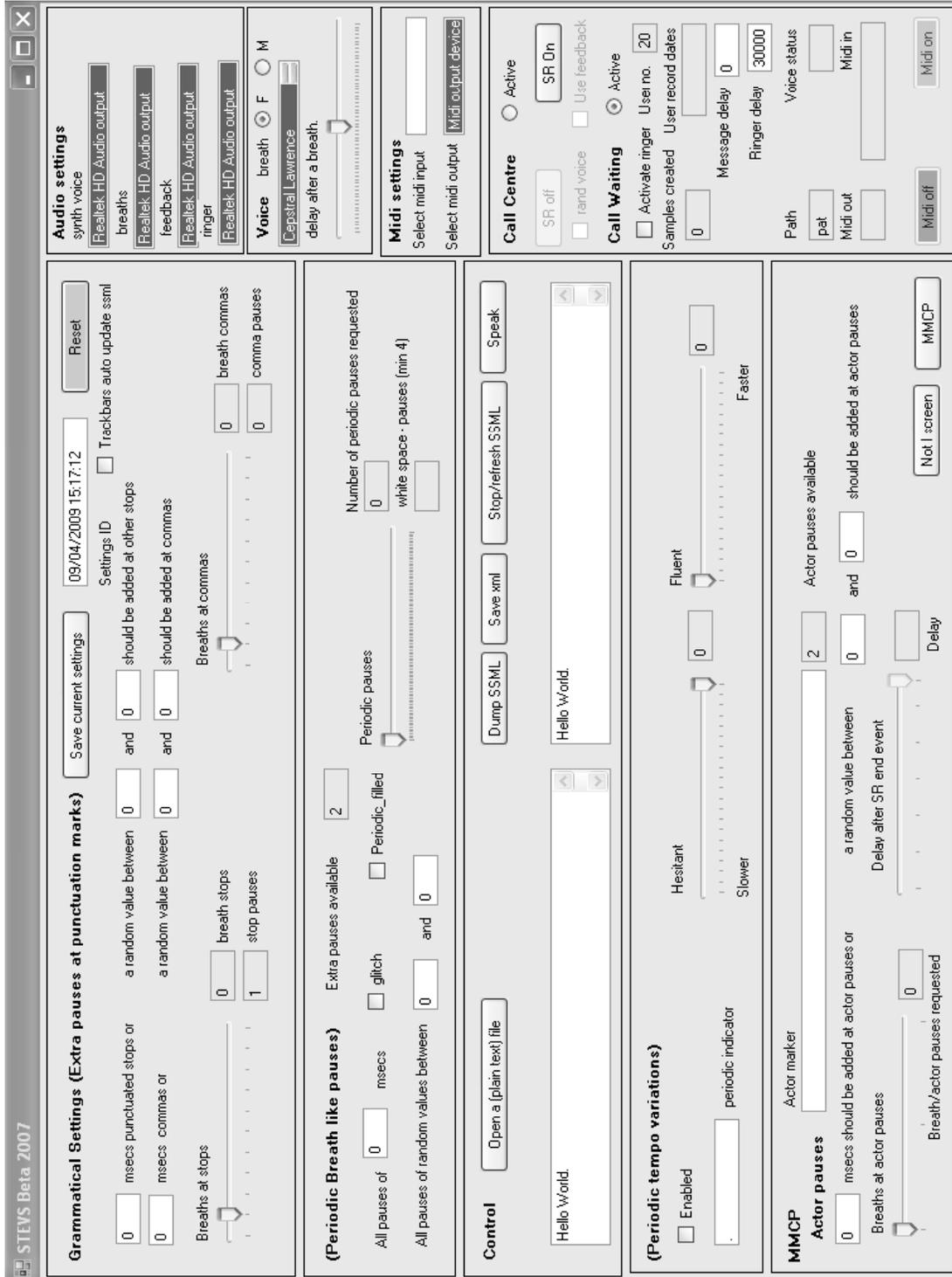


Figure 23: The interface for the PAT software tool

6.1.3 The Microsoft default settings

Careful analysis of the Microsoft TTS audio output was required to establish the reliability of the edits processed by the PAT tool. This was conducted using the PRAAT toolset (Boersma & Weenink 2008) to visually and aurally analyze waveforms representing the audio output from the Microsoft TTS pre- and post- the PAT tool processing. The concern was that the extra load on the TTS application applied by the PAT system would have audibly detrimental results. After testing, it was determined that the audible output was unaffected by the PAT system except when rendering breath audio files, when an appreciable time delay would occur.

This problem was addressed by using two audio channels, directing the playing of audio files to one and the TTS output to another. Further heuristic refinement was facilitated by providing control of the delay time, after a breath file played and before the onset of the subsequent speech.

As the choice of breath sample is determined sequentially (in a loop) or randomly, the accurate representation of audible breath in the PAT tool is constrained as shown in Table 17:

$$\text{PAT filled pause} = \left\{ \begin{array}{l} \text{Microsoft TTS default pause} \\ + \\ \text{PAT software/user defined pause} \\ + \\ \text{Random duration audible breath} \\ + \\ \text{PAT software/user defined delay} \\ \text{between 0 and 600ms.} \end{array} \right.$$

Table 17: The specification for a filled pause with breath sound in the PAT software tool

It is very important to repeat the note that, in all cases, the paralinguistic/prosodic modifiers augment the default prosodic features of the Microsoft TTS, rather than substitute for them.

This is clarified in Table 18.

Text Component or mark-up	Output from PAT software tool	Comment
.	300ms silence	As Microsoft SAPI default
. <pause 500ms>	800ms silence	PAT software tool sums the default and the SSML value
,	250ms silence	As Microsoft SAPI default
, <pause 500ms>	750ms silence	PAT software tool sums the default and the SSML value
<i>Speech rate default (i.e. no value specified)</i>	4.21 syllables per second	As Microsoft SAPI default
<i>Speech rate + 20% (80% accuracy)</i>	5.05 syllables per second	Speech rate variation proportionate to Microsoft default values
<i>Speech rate – 20% (80% accuracy)</i>	3.37 syllables per second	Speech rate variation proportionate to Microsoft default values
<i>Embedded audio file of (n) ms duration</i>	The delay time before resuming speech	The delay time before the onset of the speech chunk after play back of the audio file (user-set)

Table 18: The results of tests specifying the reliability of the PAT software tool

A further test was conducted to identify broad comparative heuristics between a human performance of a text and a reading by the Microsoft TTS system with the default settings. The text used for the test is included as Appendix D. These values would be used as a guide to the researcher for the more intuitive values implemented in the framework tests which do not make specific claims to scientific integrity (see 7.1.2 and 7.1.3).

In the test a recording of a human actor reading the text was compared with a reading by Microsoft Mary with the default settings using the PRAAT tool set for analysis.

It is interesting to note that the results recorded for speech rate variations in the 5 sentences that constituted the entirety of the speech, broadly confirm the findings of (Goldman Eisler op. cit.) in relation to the periodic speech-rate variation in human spontaneous speech compared to read speech. This gives some credence to the notion that human actors intuitively use models derived from spontaneous speech when reading. The perception of synthesised speech may be analogous to the perception of a human reading a text if the system neglects to present speech rate variation.

	Human	Micro Mary	Notes
Number of sentences	5	5	
Number of words	98	98	
Number of syllables	117	117	
Duration of reading speech + silence	36.06 secs	27.78 secs	The human performance has a longer duration
Average syllables per second	3.25 syllables per sec (in speech + silence)	4.21 syllables per sec (in speech + silence)	The human performance has the slower average speech rate. Both speak more slowly than the average recorded by Goldman Eisler for human spontaneous speech of between 4.4 and 5.9 syllables per second
Number of pauses above 200 ms	14	10	Goldman Eisler suggests that the shortest detectable pause is >250 ms. In the test the threshold was set slightly lower.
Number of words per pause	7	9.8	
Number of non-grammatical pauses	4	0	Non-grammatical pauses are not implemented in the Microsoft SAPI synthesis model
Average duration of non-grammatical pauses	0.43 secs	NA	
Total filled or unfilled pauses	11.46 secs	2.4 secs	Significantly more silence or breaths in the human performance
Average grammatical pause duration	0.81 secs	0.24 secs	The human performance has longer grammatical pauses
Fastest speech rate in one sentence	4.6 syllables per sec	NA no rate changes implemented in Microsoft Mary	Bigger speech rate variation (32.59%) compared to the average of 8 speakers recorded in Goldman Eisler's studies in spontaneous speech (21.5%) ⁵⁶
Slowest speech rate in one sentence	1.9 syllables per sec	"	
Sentence 1 duration	11.71 secs.	"	
Sentence 1 syllables	29	"	
Sentence 1 rate	2.48 syllables per sec	"	Slow
Sentence 2 duration	4.53 secs	"	
Sentence 2 syllables	18	"	
Sentence 2 rate	3.97 syllables per sec	"	Fast
Sentence 3 duration	9.36 secs	"	
Sentence 3 syllables	32	"	
Sentence 3 rate	3.42 syllables per sec	"	Slow
Sentence 4 duration	5.07 secs	"	
Sentence 4 syllables	10	"	
Sentence 4 rate	1.97 syllables per sec	"	Fast
Sentence 5 duration	6.09 secs	"	
Sentence 5 syllables	28	"	
Sentence 5 rate	4.60 syllables per sec	"	Slow

Table 19: Microsoft Mary compared to a human speaker

⁵⁶ (S.D.)/Mean x 100

6.1.4 PAT Software tool functions subject to user value assignment

The functions subject to user value assignment in the PAT Software Tool are presented in Table 20:

Variable	Theoretical range	Notes
Fixed pause-duration at full stops	0 to 10,000+ ms upper limit not tested	Goldman Eisler states that values of < 250ms are undetectable
Fixed pause-duration at commas	0 to 10,000+ ms upper limit not tested	
Random pause-duration at full stops (user-specified range)	0 to 10,000+ ms upper limit not tested	
Random pause-duration at commas (user-specified range)	0 to 10,000+ ms upper limit not tested	
Breath sounds at full stops	0 to all full stops	Random distribution of empty and filled pauses
Breath sounds at commas	0 to all commas	Random distribution of empty and filled pauses
Periodic distribution of empty pauses	0 to limit of one pause every 4 words	
Fixed-duration periodic distribution of empty pauses	0 to 10,000+ ms upper limit not tested	
Random duration periodic distribution of pauses (user specified range)	0 to 10,000+ ms upper limit not tested	
Breaths in filled periodic pauses	0 to all filled periodic pauses	Random distribution of empty and filled pauses
Actor pauses; fixed-duration-pauses at user-defined marker; paragraph break in all tests	0 to 10,000+ ms upper limit not tested	
Actor pauses; Random-duration-pauses at user-defined marker (user specified range); paragraph break in all tests	0 to 10,000+ ms upper limit not tested	
Breaths at actor pauses	0 to all filled periodic pauses	Random distribution of empty and filled pauses
Speech-rate variations at user specified marker; full stop in all tests.	Hesitant: 0 to limit of -20% Fluent: 0 to limit of +20% <i>80% accuracy in both cases</i>	

Table 20: Settings available for user modification in the PAT tool

6.1.5 Methodology for assigning values to the paralinguistic/ prosodic modifiers

The methodology applied to set the value ranges for variables in the PAT software tool was as follows:

1. Where possible, evidence would be found in the literature and translated directly into ranges that could subsequently be user-verified by items 2, 3 and 4 below.
2. A focus-group would set the range by ear, allowing for subsequent user modification or as pre-settings for tests that prohibited user modification.
3. Pilot studies would be devised to verify the range (based on the focus group) before implementation in the full scale study.
4. A full scale study based on the outcomes of item 3.

List 8: The methodology applied to set the value ranges for variables in the PAT software tool

Item 1 in List 8 was applied to produce the range of functions in the PAT software tool and the theoretical ranges shown in Table 20. In the following sections the studies to establish more precise value-ranges and implementation-settings for each of the prosodic and paralinguistic modifiers is set out.

6.1.6 Evaluating values for the prosodic/paralinguistic modifiers - focus group

Overview and goal: In order to conduct tests with values pre-assigned to the paralinguistic/prosodic modifiers implemented in the PAT software tool, a process was devised to gather a fixed set (with random components) of values based on data gathered from a focus group.

Participants: The participants were 11 undergraduates who responded to a request for volunteers. Thus the focus group was recruited on the basis of having interest in the project rather than any specific expertise. Nine were between 18 and 20 years and two were mature students of unspecified age. All were engaged in interdisciplinary arts/science courses based around digital media. Three considered themselves to be musicians (see footnote 38). None had participated in other tests in this research.

Equipment: A PC laptop running the PAT software was provided together with a pair of high quality speakers. The laptop was placed on a table in the centre of the room, with the participants seated in a semi-circle close to the speakers. The participants were unable to see the interface. The voice used was Microsoft Mary. The text chosen is included as Appendix E.

Procedure: The study was designed as an open-ended discussion lasting 30 minutes. The notion of 'liveness' was introduced, with the term 'live' as a substitute for 'liveness'. 'Liveness', a term likely to be unfamiliar to the group, was not mentioned. The objective was framed in these terms:

"We are to manipulate this chunk of synthetic speech until it sounds as 'live' as possible. To do so, we will just change the pauses and the speech-rate variations."

During the study, the participants were able to collaboratively instruct the researcher to modify the settings of the prosodic variables using the PAT tool until the objective of 'live' was met. No further explanation was provided by the researcher, but significant discussion of the problem took place within the group. The process was very animated, with many opinions expressed but consensus came easily with little disagreement.

Results: The results of the focus group study are presented in Table 21. (CD Track 45) is the unmodified synthesis of the text presented for modification to the focus group. (CD Track 46) is the final modification chosen by the focus group. The results show a significant degree of discretion being exercised in the focus group. Whereas, in the subsequent test (6.1.7), there is evidence for the users trying many modifications and arriving at a complex solution, this group was intent upon producing the required effect using only a few variables. During the process, anything that sounded excessive was generally discounted, and many of the potential features available in the software were not included in the final choice made by the focus group.

Control category	Control	Value preferred
Grammatical pauses	Set value at punctuated stops	Not selected
	Random value-range at punctuated stops	500ms – 2000ms (appended to the Microsoft SAPI default values)
	Breath sounds at stops	Not selected
	Set value at punctuated commas	Not selected
	Random value-range at punctuated commas	125ms – 350ms (appended to the Microsoft SAPI default values)
	Breath sounds at commas	Not selected
Non-grammatical pause	Set value	Not selected
	Random value	Not selected
	Breath sounds in non-grammatical pauses	Not selected
Periodic tempo variation	Hesitancy	-5% (relative to the Microsoft SAPI default values)
	Fluency	+7% (relative to the Microsoft SAPI default values)

Table 21: Results from the focus group to set values for prosodic modifiers

Caveats: There was evidence of greater caution and less playfulness in the selection of values than was shown in the subsequent web-based pilot study (1), which could result from the presence of an authority figure (the researcher). There was also a risk of social desirability factors (participants being less experimental and accepting things they privately disagree with, in order to fit in socially). The process of recruiting volunteers to the focus group was more successful than envisaged; consequently, the size of the group was more than the recommended size of 8 (Cairns & Cox 2008). Despite these caveats, the focus group agreed values for the prosodic modifiers, and were satisfied with the results.

Conclusion: Despite the availability of a number of prosodic variables, the focus group chose only four (as shown in Table 21). They rejected all variants of non-grammatical pauses, and all breath sounds. The participants chose random duration variables rather than fixed values, and this concurs with evidence of the unpredictability of pause durations in human speech (set durations tending to sound mechanical). The fact that the participants were unable to see the interface, and therefore were entirely dependent on their aural perception, may offer some

cause for greater confidence in the results than those of the subsequent test (6.1.7), the subject of the next section.

6.1.7 Evaluating values for the paralinguistic/prosodic modifiers - Web based pilot study (1)

Overview and goal: The goals of the web-based pilot study were:

1. To gather more data on user preferences for the heuristics and metrics defining the paralinguistic/prosodic modifiers.
2. To substantiate the general principle that users preferred a synthetic speech stream with some extra pauses to one without.

Although this was designed as a comparative test (a voice with no pauses judged against a voice with pauses), identified as a potential problem in 4.2, this study also offered some advantages over the focus group study and a more-conventional listening test. Individual users were able to construct their preferred performance of a preset text in isolation, without an authority figure present (in the room), and submit their choice online for analysis by the researcher. They were also able to explore, in their own time, the full potential of each variable, going back and forth, selecting and deselecting edits, until satisfied with the results. This provides a more personal and considered response; one less likely to be influenced by other participants than a focus group study.

The study allowed the participant to experience the combinatorial perception of multiple pause types, “P” (“PG” for grammatical pauses, “PB” for breath-like pauses, “PR” for random pauses or none), and durations, “D” (“DR” random, “DS” short, “DL” long or none). For the purposes of analysis, each pause-type and duration pair would be treated as a discreet data point, and therefore the results would demonstrate the effectiveness of each pair composed of “P D” in isolation and not when combined with others⁵⁷.

⁵⁷ An alternative method of analysis could be considered if there was a very large group of participants. Each participant would submit a preferred edit, containing 3 sets of from 1 to 4 P and D pairs (P1 and (DR or DS or DL or none)), (P2 and (DR or DS or DL or none)), (P3 and (DR or DS or DL or none)). The patterns of pairs could be compared until a significant preference trend emerged. However, the probability of a significant trend emerging with a small group of participants would be very low as 64 editing combinations are possible for each participant submission.

Participants: Participants were invited via e-mail communication and word-of-mouth. Forty-nine participants submitted results. One submitted twice (the second submission was flagged as ‘worst’ by the user and has been discounted; thus, 48 individual participants submitted valid results. To optimise uptake, and to keep the test short, no personal data, other than the participant’s e-mail address and whether they played a musical instrument (see footnote 38), was gathered. Twenty-five participants played musical instruments. From an examination of the e-mail addresses it is possible to conclude that the composition of the group was broadly comparable to the focus group; it was composed of 40 undergraduate students, but also included 8 adult friends or colleagues known to the researcher but with no knowledge of this research.

Equipment: A web application was built with a subset of features derived from the PAT software tool and one feature not included in the PAT software tool, providing three randomly positioned (ungrammatical) pauses over the duration of the text. The voice used was Microsoft Mary. The audible effect of the prosodic modifications was tested using the PRAAT audio analysis software (Boersma & Weenink 2008) and found to be identical to those of the PAT software tool. The following technical differences should be noted:

- The application was written in Javascript for Microsoft Internet Explorer.
- The runtime application environment was the Microsoft Active X Control Speech Add-in, derived from the Microsoft Speech SDK used for the PAT software tool.
- The prosodic instructions were encoded using Speech Application Language Tags (SALT) and a version of SSML designed for the Active X Control.

The interface significantly simplified the choice of prosodic modifications available to the user; however, some of the same broad categories of pause durations and distributions were made available, plus one new category.

The pauses were in three categories, each category with four options, thus:

1. Grammatical pauses at full stops:
 - a. None
 - b. Random-length
 - c. Short

- d. Long
- 2. Breath-like pauses (non-grammatical pauses after approximately 11 syllables):
 - a. None
 - b. Random-length
 - c. Short
 - d. Long
- 3. Randomly-placed (non grammatical pauses):
 - a. None
 - b. Random-length
 - c. Short
 - d. Long

The setting ranges applied to the prosodic and paralinguistic modifiers in the web-based pilot study (1) are shown in Table 22. The setting for breath-like-pauses was set by the researcher and based on an approximate simulation of the breath pauses taken after 10 syllables as specified in 2.5.3.

	Random-length	Long	Short
Pause at full stops. Extra silence would be added at a grammatical full stop.	0 – 1000 ms	500 ms	125 ms
Breath-like pauses. Extra pauses would be added after every 7 words.	0 – 1000 ms	500 ms	125 ms
3 randomly-located pauses would be distributed across the text.	0 – 1000 ms	500 ms	125 ms

Table 22: Setting ranges applied to the prosodic and paralinguistic modifiers

Procedure: The same short passage used in the focus group was spoken by Microsoft Mary (included as Appendix F), and the user was able to modify the speech by adding pauses.

The user was expected to read the instructions before starting the test, as follows:

“Thank you for participating in this experiment. Our objective is to find out whether a synthetic voice can be made to sound more 'alive' just by adding extra pauses to the speech stream.

To do this you will be given some tools to choose from and combine to add pauses. When you are satisfied with the performance of the voice you can comment on it and then submit the results to the researchers for analysis. In the meantime here is the speech. You can see it is not very long.”

Figure 24 shows the interface for the web-based pilot study (1):

Play with the radio buttons on the left to change the way the text is spoken. Some changes are subtle others are very obvious. You can also combine edits. Some choices would add the same pause type on top of the same pause type. This would make the results from the experiment hard to quantify. The system will prevent you making those choices. Otherwise feel free to add pauses to try to make the most 'live' sounding voice. You can decide what 'live' means. When you have something you like, comment and submit. Please don't submit your least successful edits as that is pretty easy to do and does not help the experiment.

E-mail address: Do you play a musical instrument? No Yes

But on the other hand, despite that, I will only ever be what people want me to be. I'll be a nurse or a soldier or a runaway bride or a grumpy woman in a tea shop. But I can never be me. So I can't do what you want me to. You're asking too much of me, Adam.

Add extra grammatical pauses at full stops.

No extra
 Random duration
 Long
 Short

Add breath-like pauses

No breaths
 Random duration
 Long
 Short

Add randomly positioned pauses

No random
 Random duration
 Long
 Short

System messages Click 'stop' when you have heard enough to judge your edit

Your comments on this edit

My best
 Fair
 My worst

Figure 24: Interface for the web-based pilot study (1)

The interface presented the user with a series of radio buttons, allowing for the selection of a range of pause durations and distributions. Only one choice could be made per set of edits, but three sets of edits could be applied in layers. After each edit was made, the system would remind the user to listen to the modifications made. All the edits could be cleared by selecting 'Clear Edits'. After creating their preferred edit, the user could submit a single choice for analysis, and add a comment and a rating between 'My best', 'Fair' and 'My Worst'. They were encouraged not to submit their "Worst"⁵⁸, and the system required them to include an e-mail address, a comment and a rating before submission was accepted.

Results: The results of the web-based pilot study are presented in Table 23. A sample of an atypical (for clarity) extreme editing choice is included as:

- (CD Track 47) The original text with no edits (for reference).
- (CD Track 48) Long grammatical pauses.
- (CD Track 49) Long grammatical pauses and long breath-like pauses.
- (CD Track 50) Long grammatical pauses, long breath-like pauses, long random pauses.

Pause Type	None (edits)	Random-duration (edits) R	Short-duration (edits) S	Long-duration (edits) L	Total edits executed R+S+L	P =
Grammatical distribution	10	11	17	10	38/48	0.418
Breath-like distribution	7	15	14	12	41/48	0.367
Random distribution	23	8	12	5	25/48	0.001
Totals	40	34	43	27	104/144	

Table 23: Results of web-based pilot study (1). The P column shows the p-value for a chi-square test of this data as reported by Excel software.

⁵⁸ In retrospect this part of the test procedure was an unnecessary complication that did not produce any additional data of value. The intention was to allow users to comment on the edit without typing, however it could have led to users assuming that they were required to submit a set of graded responses. The decision to collapse all 'fair' and 'good' submissions into one (for analysis purposes) should be noted as a caveat in this study. No valid 'worst' submissions were made.

Table 23 shows the number of users selecting the different edit-types and submitting them as their ‘best’ or ‘fair’ editing choice (see Footnote 58). From this the reader is able to see that certain edits were more popular than others; for example, short-duration, grammatically-distributed edits were chosen seventeen times, whereas long duration randomly distributed edits were only chosen five times. In Figure 25 the graphs show the frequency of edit-types for each of the three pause-categories. The Y-axis represents the number of users submitting an edit as “best” or “fair”. The X-axis shows the four pause-duration-categories selected by users: none, random-duration, long-duration, and short-duration.

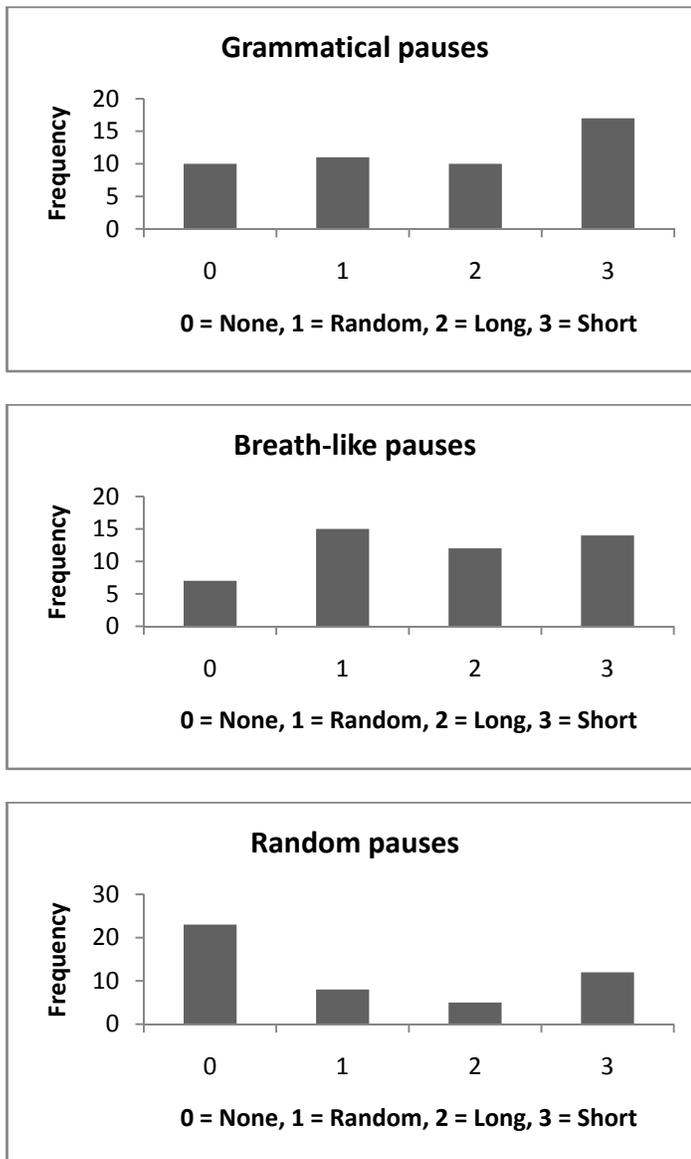


Figure 25: The frequency of edits for each of the three categories of pauses

The p-value for a chi-square test of this data as reported by Excel software shows that the results are only statistically significant in relation to the user preference against embedding extra random pauses.

The results for the duration of grammatically-distributed pauses and pauses with a breath-like distribution are not statistically significant. However in order to achieve the second goal of this study: 'To substantiate the general principle that users preferred a synthetic speech stream with some extra pauses to one without' if the results for each category of executed pauses types are summed (labeled 'any edits' in Table 24) and compared against the 'no edits', a different result emerges.

Pause Type	No (edits)	Any duration (edits)	P =
Grammatical distribution	10	38	0.000
Breath- like distribution	7	41	0.000
Random Distribution	23	25	0.773
Totals	40	104	

Table 24: Results of any pause edits compared to no pause edits. The P column shows the p-value for a chi-square test of this data as reported by Excel software.

Table 24 shows a statistically significant user preference for a grammatical and breath-like distribution of pauses, of any of the durations available in the test. However, no statistical significance can be shown for the user preference for random-distribution pauses of any of the durations available in the test.

Conclusion: The study ran during the period that the focus group met; hence some results from the focus group came too late to influence the design of the web-based pilot study. The disparity in the potential variables in this test from those present for the focus group (in particular the omission of speech-rate variations and comma pauses, in this study) may reduce the value of these results. On the other hand, to have included the complete set of variables offered to the focus group might have made the test too complicated for online participants,

resulting in a low up-take for the test and thus weaker results. The caveat presented by collapsing the results for the 'best' and 'fair' edits into one set is qualified in footnote 58.

The contradictory result for the random pauses when subject to two different analyses is problematic. When considered in relation to the focus group findings in which breath-like pauses (that have a similar auditory effect to random pauses) were rejected, it seems probable that the first analysis, which intimates that the participants reject the use of random pauses, is more reliable than the second, which in any case is not statistically significant.

The most significant finding was that users did in fact prefer pauses to no pauses, but there is a possibility that an inclination to 'play' with the software and a desire to please the researcher by making lots of modifications may have influenced the results. The analysis of the results must also be subject to caution. While it is true that summing the results from all the pause-duration variables in each category of pause type produces a conclusive outcome for grammatical and breath-like pauses, it can be argued that this distorts the results by presenting an outcome based on a choice of four variables as if it were based on a choice of two.

Allowing for these potential caveats, this study revealed a high probability that users prefer a synthetic speech stream with extra pauses to one without them. It did not suggest any more-exact metrics for the durations of the pauses, for which no statistically significant data was gathered.

6.1.8 Evaluating values for the paralinguistic/prosodic modifiers - Web based pilot study (2)

Overview and goal: This test was designed to provide more comprehensive evidence leading to the correct parameterisation of the variables in the PAT software tool. By presenting a more extensive set of pre-rendered speech chunks as well as three different voices (including two different synthetic voices, one human voice and sound effects), the hope was that some additional evidence for appropriate generalisable metrics for the paralinguistic/prosodic modifiers could emerge. In this study the metrics and effects applied to the prosodic variables were set to more extreme and experimental values than those in the focus group or web based pilot study (1). The intention was to address the issues raised by (4.2 Hearing liveliness test) in which the results indicated that liveliness rendered as tones without linguistic content may

need to be 'over-acted' to be detectable by users. Other metrics were based on the heuristics posited by the human performance documented in Table 19 as well as settings based on interpretations of the sources and perspectives specified in 2.5.

A pilot study with seven participants was conducted in advance of a full study. As the stated intention was a more comprehensive view of paralinguistic/prosodic modifiers applied to different voices, there was some concern that the participants would have difficulty seeing past the many confounding variables. On the other hand, if the confounding variables did not overwhelm the participants' discretionary powers to tease out the significant prosodic variation intended by the researcher, then the study would strengthen the findings from the previous two studies. In order to assess the validity of the study, two identical pairs of recordings were positioned at different points in the test sequence. Inconsistent scores for the recordings in each pair would show the test to be flawed. In this study, the notion of 'liveness' was explained as the difference between the voice 'reading' or 'speaking', summed up as the 'best voice actor'.

Participants: Six of the seven volunteer participants were final year undergraduate students engaged in interdisciplinary arts/science courses based around digital media. Three played a musical instrument. All 6 undergraduates were in their 20s. One participant was a middle-aged lecturer. None had participated in other studies related to this research. A draw for a bottle of wine was offered as an enhancement.

Equipment: The test was presented as web-based questionnaire in Microsoft Internet Explorer. The participants played a set of audio files sequentially. The audio files were produced by the PAT software tool, rendered in Macromedia Shockwave format, and played over headphones. The audio files are included on the CD Track 51 to Track 64. After each group of audio files was played, the participant was required to rank the recordings. In tests 1, 2 and 4, this involved choosing the worst and the best voice actor (from the sets of 3, 3 and 2 respectively). In test 3 this involved ranking the 6 performances from 1 to 6, with '1' as the best and '6' as the worst. After playing all the recordings the participant submitted the results for storage in a Microsoft Access database. The system allowed participants to submit incomplete or null results. The application was written using Macromedia Cold Fusion Mark-Up Language CFML.

Procedure: Each participant in the pilot study underwent the test while overlooked by the researcher in an office at a UK university. No explanation was provided other than the written instructions on the screen. By observing, the researcher was able to ensure that each candidate followed the same sequence. A brief level-test was conducted to ensure that the headphone volume-level suited the participant. Figure 26 shows the instruction screen. The text used is included as Appendix G.

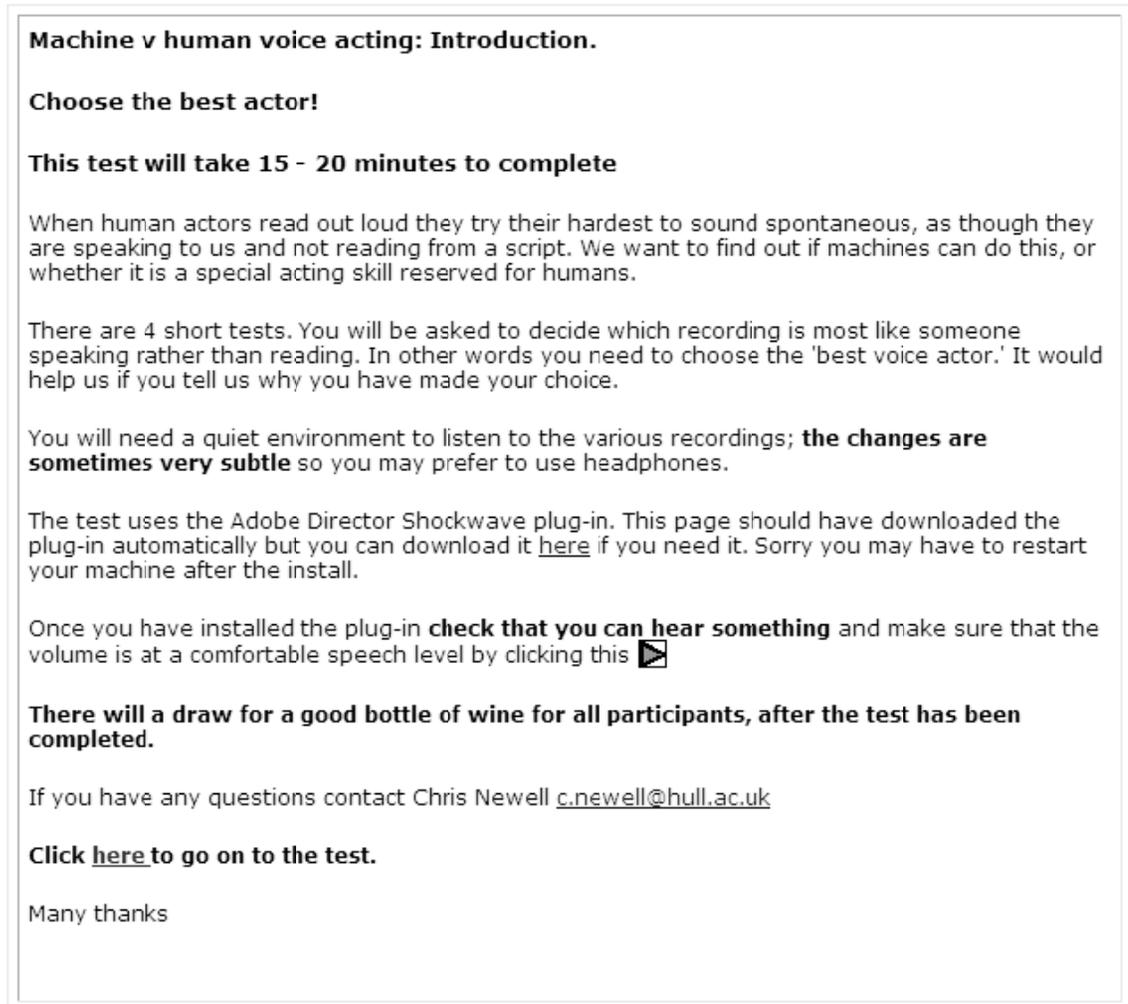


Figure 26: Instruction screen for web based pilot study (2)

After reading the instructions, the participant would be presented with a single scrolling screen showing the introduction (Figure 27), tests 1-3 (Figure 28 to Figure 30) test 4 and the summary and submission dialogue (Figure 31).

INTRODUCTION		
Please wait until the page has a  next to every sample.		
There are a total of 5mb of sounds to download and it may take a few minutes for the sound files to load. We have to load them all in advance as an interruption to the play back would spoil the test.		
<ol style="list-style-type: none"> 1. Please go through each section; 1, 2 3 and 4 in turn. Scroll down if you cannot see all the sections. 2. Play each sound one after the other in numerical order by clicking on the play  button. The button will turn red as the sound starts to playback. 3. In Test 3 you may play each sample as many times as you like. 4. When you have completed all 4 sections submit your answers. 5. If something much more interesting interrupts your test, submit the incomplete test anyway. 		
We need to know just a few things about you to make sure the data we gather is as fair as possible. The data will be retained by the researchers but your entry will be anonymous.		
Please enter you e-mail address in order that we can notify you if you win the wine. If you prefer not to enter the draw, leave this field and the next row blank.	E-Mail	<input type="text"/>
Which wine would you prefer if you win?	<input type="radio"/> Red	<input type="radio"/> White
Your gender?	<input type="radio"/> Male	<input type="radio"/> Female
Do you play a musical instrument?	<input type="radio"/> Yes	<input type="radio"/> No
Are you an actor or voice performer?	<input type="radio"/> Yes	<input type="radio"/> No
Do you use a speech synthesiser more than once a month?	<input type="radio"/> Yes	<input type="radio"/> No

Figure 27: Introductory screen for web based pilot study (2)

TEST 1				
Choose which sample sounds as though someone is speaking rather than reading (best voice actor) then choose which sample sounds as though someone is reading rather than speaking (worst voice actor).				
Sample	Play	Speaking (best voice actor)	Reading (worst voice actor)	Please comment on your decision. (max 255 characters)
Sample 1a	 Please listen once only.	<input type="radio"/>	<input type="radio"/>	<input type="text"/>
Sample 1b	 Please listen once only.	<input type="radio"/>	<input type="radio"/>	<input type="text"/>
Sample 1c	 Please listen once only.	<input type="radio"/>	<input type="radio"/>	<input type="text"/>

Figure 28: Test 1 for web based pilot study (2)

TEST 2				
Choose which sample sounds as though someone is speaking rather than reading (best voice actor) then choose which sample sounds as though someone is reading rather than speaking (worst voice actor).				
Sample	Play	Speaking (best voice actor)	Reading (worst voice actor)	Please comment on your decision. (max 255 characters)
Sample 2a	<input checked="" type="checkbox"/> Please listen once only.	<input type="radio"/>	<input type="radio"/>	<input type="text"/>
Sample 2b	<input checked="" type="checkbox"/> Please listen once only.	<input type="radio"/>	<input type="radio"/>	<input type="text"/>
Sample 2c	<input checked="" type="checkbox"/> Please listen once only.	<input type="radio"/>	<input type="radio"/>	<input type="text"/>

Figure 29: Test 2 for web based pilot study (2)

TEST 3		
Put these samples in order. The one that sounds most as though someone is speaking rather than reading (best voice actor) gets a score of 6		
If you want to give several samples the same value i.e. 'Sample 5 was most like someone is speaking rather than reading (best voice actor) but the rest were all about the same', then just give a score of 6 to sample 5 and leave the rest blank. You can also give the same score to several samples.		
Important: You can play these samples as many times as you like, until you are really sure of your answer.		
Sample	Play	Score
Sample 3a	<input checked="" type="checkbox"/>	<input type="text"/> 6 is speaking rather than reading (best voice actor). 1 is reading rather than speaking (worst voice actor).
Sample 3b	<input checked="" type="checkbox"/>	<input type="text"/> 6 is speaking rather than reading (best voice actor). 1 is reading rather than speaking (worst voice actor).
Sample 3c	<input checked="" type="checkbox"/>	<input type="text"/> 6 is speaking rather than reading (best voice actor). 1 is reading rather than speaking (worst voice actor).
Sample 3d	<input checked="" type="checkbox"/>	<input type="text"/> 6 is speaking rather than reading (best voice actor). 1 is reading rather than speaking (worst voice actor).
Sample 3e	<input checked="" type="checkbox"/>	<input type="text"/> 6 is speaking rather than reading (best voice actor). 1 is reading rather than speaking (worst voice actor).
Sample 3f	<input checked="" type="checkbox"/>	<input type="text"/> 6 is speaking rather than reading (best voice actor). 1 is reading rather than speaking (worst voice actor).
Please comment on your decision. (max 255 characters)		<input type="text"/>

Figure 30: Test 3 for web based pilot study (2)

TEST 4				
Choose which sample sounds as though someone is speaking rather than reading (best voice actor) then choose which sample sounds as though someone is reading rather than speaking (worst voice actor).				
Sample	Play	Speaking (best voice actor)	Reading (worst voice actor)	Please comment on your decision. (max 255 characters)
Sample 4a	 Please listen once only.	<input type="radio"/>	<input type="radio"/>	<input type="text"/>
Sample 4b	 Please listen once only.	<input type="radio"/>	<input type="radio"/>	<input type="text"/>

TEST SUMMARY	
Please add any general comments on any of the samples on the page. (max 255 characters)	<input type="text"/>
Thank you very much.	<div style="display: flex; justify-content: space-between; align-items: center;"> <div>Please go ahead and submit your answers.</div> <div><input type="button" value="Submit"/></div> </div>
Human voice and script fragment by Stuart Andrews. S.Andrews@hull.ac.uk	

Figure 31: Test 4 and summary dialogue box for web based pilot study (2)

The recordings participants heard are listed in Table 25.

Test	Recording specification	Notes	CD reference
1a	Recording of human actor; fixed-duration (2000 ms) pauses inserted at each full stop.	Human	(CD Track 51)
1b	Recording of human actor; no extra pauses	Human; the unmodified performance	(CD Track 52)
1c	Recording of human actor; random-duration (between 500 and 2000 ms) pauses inserted at each full stop	Human	(CD Track 53)
2a	Recording of synthetic voice; grammatical pauses (durations duplicated from the human recording)	Microsoft Sam Same as 3f	(CD Track 54)
2b	Recording of synthetic voice; random-duration (between 500 and 2000 ms) pauses at full stops aligned to regular bar lines. ⁵⁹	Microsoft Sam	(CD Track 55)
2c	Recording of synthetic voice; no extra pauses	Microsoft Sam Same as 3d	(CD Track 56)
3a	Recording of synthetic voice; random-duration (between 500 and 2000 ms) pauses at full stops; background sound (waves)	Microsoft Sam	(CD Track 57)
3b	Recording of synthetic voice; random-duration (between 500 and 2000 ms) pauses at full stops	Microsoft Sam	(CD Track 58)
3c	Recording of synthetic voice; random-duration (between 500 and 2000 ms) pauses at full stops; background sound (a clock ticking)	Microsoft Sam	(CD Track 59)
3d	Recording of synthetic voice; no extra pauses	Microsoft Sam Same as 2c	(CD Track 60)
3e	Recording of synthetic voice; random-duration (between 500 and 2000 ms) breath-filled pauses at full stops.	Microsoft Sam	(CD Track 61)
3f	Recording of synthetic voice; grammatical pauses (durations duplicated from the human recording)	Microsoft Sam Same as 2a	(CD Track 62)
4a	Recording of alternative synthetic voice	Loquendo Simon	(CD Track 63)
4b	Recording of synthetic voice; random-duration (between 500 and 2000 ms) breath-filled pauses at full stops; background sounds (waves and clock)	Microsoft Sam	(CD Track 64)

Table 25: Description of audio recordings for web based pilot study (2)

⁵⁹ For this recording the onset of each utterance was aligned to the beginning of an arbitrary but consistent length bar. Thus the intention was to create a subliminal pulse as described by both Stanislavski (see 2.5.2) and Wennerstrom (see 2.4)

Results: Four out of 7 participants showed inconsistencies in evaluations of the same recordings in different positions in the test. One participant evaluated the same recordings as both the best and the worst in different tests. Although this may indicate complex interactions occurring between the perceptions of recordings in different contexts, a more realistic conclusion is that the test is flawed. A summary of the results are shown in Table 26 and Table 27. No results of statistical significance emerged from this pilot study.

Voice	Recording	Total 'best actor'	Total 'worst actor'	Total null	Total (best – worst)	P =
Human	1a	0	5	2	-5	0.06
Human	1b	2	1	4	1	0.36
Human	1c	5	0	2	5	0.06
Microsoft Sam	2a	2	3	2	-1	0.86
Microsoft Sam	2b	1	1	5	0	0.10
Microsoft Sam	2c	3	2	2	1	0.86
Loquendo Simon	4a	3	2	2	1	0.86
Microsoft Sam	4b	3	4	0	-1	0.15

Table 26: Results of tests 1, 2 and 4 for web based pilot study (2). The P column shows the p-value for a chi-square test of this data as reported by Excel software.

Recording	Total score	Ranking
3a	20	4
3b	17	5
3c	21	3
3d	25	2
3e	14	6
3f	27	1
	P= 0.34	

Table 27: Results of test 3 for web based pilot study (2). P shows the p-value for a chi-square test of this data as reported by Excel software.

Conclusion: The pilot study showed that the effectiveness of selected paralinguistic/prosodic modifiers could not be demonstrated in a test in which a large range of additional variables were presented to the user. It showed that other factors, such as voice types and background sounds, would overwhelm the participant's discriminatory powers to detect the potential enhancements provided by the chosen paralinguistic/prosodic modifiers. The study was poorly designed and it allowed the researcher to indulge in too many speculative compositional renderings of the prosodic/paralinguistic modifiers better suited to an exhibition than an experiment. This realisation directed the researcher toward a deeper understanding of the complex interrelated variables that operated in any interaction between a human and a synthetic voice, and toward the notion of a more holistic framework in which these complexities could be examined. The framework would also provide the opportunity to try out the more speculative renderings of the prosodic/paralinguistic modifiers (see 7.1.2). This setback also led to the design of the user survey (subject of the next section) in which the confounding variables are reduced and the result are clearer.

6.1.9 Evaluating values for the paralinguistic/prosodic modifiers - User survey

Overview and goal: This test was designed to support or qualify the findings from the focus groups and web based pilot study (1). Although the principle of scoring a range of audio recordings adopted in web based pilot study (2) was used, this was a much simpler test using only one synthetic voice and six recordings. The participants selected two recordings - one as the worst, and one as the best - but could provide comments on all if they wished. The six recordings included one synthetic voice recording with the grammatical pauses durations replicated from a reading by a human actor and another using the focus group settings. One recording was obviously unrealistic with inappropriate pauses and quite semantically disturbing. The differences between the others were subtle and demonstrated some of the features rejected by the focus group.

Participants: The test was conducted with a single group of 52 participants completing a paper based test. Forty-nine participants returned valid questionnaires. The participants were principally middle aged, but there were younger people and some of retirement age. Twenty four played musical instruments and one used computer generated speech more than once a month. The participants were gathered in the evening in a village hall in North Yorkshire, UK.

Age data was not collected, although children less than 13 years were asked to indicate this on the questionnaire (none did). The primary purpose of the gathering was a village 'Blues' musical concert (which may account for the high proportion of musicians in the sample). No participant had participated in any of the other studies in this research. A draw for a bottle of wine was offered as an incentive.

Equipment: The recordings were produced as Macromedia Shockwave files using the PAT software tool, and were played from a laptop over hi-fidelity speakers. Each recording was as specified in Table 28 and presented as (CD Track 67 to CD Track 72). The settings were based on heuristics gathered from the previous studies. The voice used was Microsoft Sam. The text used is presented in Appendix I.

Procedure: The participants were provided with a paper based questionnaire and a pen. They were seated 'cabaret-style' (in small groups around individual tables), but were asked not to confer. No instructions other than those printed on the questionnaire were provided (Figure 32). The study was conducted largely in silence, although some participants found that a challenge. A sample of a completed questionnaire is shown in Figure 33⁶⁰.

⁶⁰ The date shown on the questionnaire was the original date planned for the test. The test actually took place on January 19th 2008 and the questionnaires were printed with the incorrect date.

PAT Test 7th Dec 2007

PAT – Test 7th Dec 2007

Instructions:

Before commencing the test please read the instructions carefully.

This test takes approximately 20 minutes to complete.

1. When human actors read out loud they try their hardest to sound spontaneous, as though they are speaking to us and not reading from a script. We want to find out if machines can do this, or whether it is a special acting skill reserved for humans.
2. Each sample lasts about 40 seconds.
3. After hearing sample 6, all 6 samples will be repeated once.
4. There are 6 different samples. They all use the same text and they are all have the same voice quality. You will be asked to choose one 'best voice actor' and one 'worst voice actor.' Do not complete the judgment column until you have heard all the samples once. You can use the *Comment* column to remind yourself of your first impression, but we would also like to examine your comments, so it would be good if you could write a comment for every sample. Simple, brief but legible notes will suffice.
5. For the best actor write 'best' in the judgment column. For the worst actor write 'worst' in the judgment column.
6. It would help us if you tell us what you think of each sample in the comment section. You may wish to point out features that stand out as particularly poor or comment on positive features.
7. You will need a quiet environment to listen to the various recordings; the changes are sometimes very subtle.

Section 1: Subject details			
Your gender?	Male	<input checked="" type="checkbox"/>	Female
Do you play a musical instrument?	Yes	<input checked="" type="checkbox"/>	No
Are you an actor or performer?	Yes	<input checked="" type="checkbox"/>	No
Do you use computer generated speech more than once a month? (For example: a screen reader).	Yes	<input type="checkbox"/>	No <input checked="" type="checkbox"/>
Section 2: Prize draw			
If you would like to enter a draw for a bottle of wine please enter your e-mail address.	Email	<input type="text"/>	

Figure 32: Instruction page for user survey

PAT Test 7th Dec 2007

Section3: Samples test		
	Comment: (Please comment on all the samples)	Judgment: (One best One worst)
Sample 1	All wrong. Misplaced inflection. Panicky/childlike. Babble.	Sucked like pop Idol. <u>The</u> Worst.
Sample 2	Questioning. Like Australian. Disappointed.	
Sample 3	Sounds unsure of what it is saying.	
Sample 4	Relatively Natural.	Best.
Sample 5	Strident. Guttural almost. Irritating	
Sample 6	Dunno. Babbling a bit	

Figure 33: Sample answer page for user survey

Recording	Total 'best actor'	Total 'worst actor'	P =	('best' minus 'worst' actor)
1. Averaged grammatical pause-values (based on a human performance). (CD Track 67)	8	8	0.013	0
2. Focus group settings (see 6.1.6) (CD Track 68)	10	1	0.002	+9
3. Human-length grammatical pauses (based on a human performance). (CD Track 69)	20	0	0.000	+20
4. No changes made. (CD Track 70)	5	8	0.073	-3
5. Periodic breath-like pauses of 470 ms (based on a human average duration for the speech chunk) distributed evenly. (CD Track 71)	0	30	0.000	-30
6. Fluent and hesitant tempo variations of 10% each way. (CD Track 72)	6	2	0.168	+4
Valid responses	49	49		

Table 28: Results of the user survey. The P column shows the p-value for a chi-square test of this data as reported by Excel software.

Results: In table 29 the χ^2 test showed statistical significance for the results for recordings 1, 2, 3 and 5. Recording 1 had as many positive as negative responses. The results for recording 3 appear to be the best, with recording 2 only half as well liked by the participants. Recording 5 is the least well liked.⁶¹

Conclusion: The endorsement by this group of participants of recording 3, human-length grammatical pauses (based on a human performance) may show sensitivity among the subjects to the human capacity for setting appropriate pause durations. Recording two, using focus-

⁶¹ A detailed text analysis of all 294 written comments may provide scope for additional qualitative research but the scope of the thesis prevents any further work on this data at this stage. Of particular interest is the use of terms in comments such as 'natural' (11), 'emotion' (10) demonstrating some participant's tendency to measure the synthetic voice against human verisimilitude. It is also of interest that the term speech (12) and voice (14) seemed to be used interchangeably. The term 'pause' occurred (36) times although the users were not informed that this was the principal variable. The term 'clear' occurred (38) times suggesting that despite industry claims, users are still troubled by issues of clarity or using it as a qualitative measure.

group settings, received only one 'worst actor' rating and was the second most popular choice. This result confirmed the positive endorsement of these settings provided by the focus group, and was encouraging. The unmodified recording was only 25% as popular as the 'best' modified recording and was in 5th place with only one recording (the most semantically disturbing) scoring lower. This study provided evidence for some confidence that the settings specified by the focus group were generalisable to other user groups in a different test context. This test provided additional evidence that, despite the variability intrinsic in setting filtered random length pause durations (within ranges set by the focus group) they could provide a suitable second choice after the intelligent encoding of specified pause durations according to a human performance.

6.1.10 Evaluating the perception of 'liveness' - Web-based pilot study (3)

Overview and goal: This study builds on the positive outcome of the focus group and user survey, and attempts to sidestep the issue of identifying the perception of 'liveness' by examining more closely the degree of engagement a user may demonstrate that may betray the presence of the perception of liveness. Using this method, the researcher would have additional evidence of the effectiveness of the paralinguistic/prosodic modifier settings derived from the two earlier studies.

This study attempted to analyze different levels of perceptive enhancement that could be stimulated by appropriate paralinguistic/prosodic modifier settings. This was done by asking questions that do not directly relate to 'liveness', but would test other issues, such as comprehension and engagement, that may indicate liveness's perception. The basis for this assumption was that something that had 'liveness' would be more engaging and therefore command more attention and be more memorable. This methodology is much more closely aligned to the methodologies used in more conventional user studies on synthetic speech such as those documented by Nass and Brave (Nass & Brave op. cit.) and Whalen et al. (Whalen, Hoequist & Sheffert op. cit.). The test presented two performances of a longer text. One performance would be unmodified; the other performance would apply modifications to the speech stream in line with the findings from Web based pilot study (1) (see 6.1.7) the focus group (see 6.1.6) and the user survey (see 6.1.9). The subjects would listen to one or the other

performance, and then answer questions. The null hypothesis was that the extra pauses would make no difference to engagement or comprehension.

Participants: Participants were invited by e-mail or word of mouth. Thirty-five participants took part in the study. 17 participants heard the unmodified voice; 18 heard the modified voice. They consisted of friends and relatives of the researcher, and students with an age range from late teens to middle age. None had taken part in the other studies. Twenty-one played a musical instrument. A draw for a bottle of wine was offered as an incentive.

Equipment: The test was presented as web-based questionnaire in Microsoft Internet Explorer. A single audio file of one or other of the two versions, rendered in Macromedia Shockwave format, could be played through headphones or speakers by the participant. After the audio file was played, the participant was required to answer an online questionnaire. On completion, the participant submitted the results for storage in a Microsoft Access database. The application was written in Macromedia Cold Fusion Mark-Up Language (CFML).

Procedure: The test was unsupervised, with each participant undergoing the test in the environment of their choice. No explanation was provided other than the written instructions on the screen. A short level-test was offered to ensure that the volume level suited the participant. Figure 34 shows the introduction screen.

Machine v human voice acting: Introduction to PPP test.

Choose the best actor!

This test will take 15 minutes to complete

When human actors read out loud they try their hardest to sound spontaneous, as though they are speaking to us and not reading from a script. We want to find out if machines can do this, or whether it is a special acting skill reserved for humans.

There is one short listening test after which you will be asked to answer 10 questions.

You will need a quiet environment to listen to the recording; **the changes are sometimes very subtle** so you may prefer to use headphones.

The test uses the Adobe Director Shockwave plug-in. This page should have downloaded the plug-in automatically but you can download it [here](#) if you need it. Sorry you may have to restart your machine after the install.

Once you have installed the plug-in **check that you can hear something** and make sure that the volume is at a comfortable speech level by clicking this 

There will a draw for a good bottle of wine for all participants, after the test has been completed.

If you have any questions contact Chris Newell c.newell@hull.ac.uk

- When you click the button below the browser will ask you if this window can be closed.
- It will open a new one.
- You should close this window which may have dropped behind the new window.

Many thanks.

Figure 34: Introduction screen for web-based pilot study (3)

On the next screen (Figure 35), the participant was invited to listen to the audio sample. Playback lasted either 2 minutes 42 seconds or 3 minutes 30 seconds, depending on which version was played (unmodified or modified). The voice used was Microsoft Mary. The text for this test is presented as Appendix H.

INTRODUCTION		
Please wait until the page has a  next to the sample.		
There is a total of 5mb to download and it may take a few minutes for the sound file to load. We have to load it in advance as an interruption to the play back would spoil the test.		
<ol style="list-style-type: none"> 1. Play the sound by clicking on the play  button. The button will turn red as the sound starts to playback. 2. When the playback finishes go on to the next page. 		
Sample	Play (please wait for green play button below)	
59 <input type="text"/>	 Please listen once only.	Next page

Figure 35: 'Play' screen for web-based pilot study (3)

The participant was invited to complete personal details (Figure 36).

TEST		
Please go through the questions below in order. When you have answered all the questions submit your answers by clicking the submit button.		
We need to know just a few things about you to make sure the data we gather is as fair as possible. The data will be retained by the researchers but your entry will be anonymous.		
If you cannot answer a question leave it blank.		
Please enter you e-mail address in order that we can notify you if you win the wine. If you prefer not to enter the draw, leave this field and the next row blank.	E-Mail	<input type="text"/>
Which wine would you prefer if you win?	<input type="radio"/> Red	<input type="radio"/> White
Your gender?	<input type="radio"/> Male	<input type="radio"/> Female
Do you play a musical instrument?	<input type="radio"/> Yes	<input type="radio"/> No
Are you an actor or voice performer?	<input type="radio"/> Yes	<input type="radio"/> No
Do you use a speech synthesiser more than once a month?	<input type="radio"/> Yes	<input type="radio"/> No

Figure 36: Participant details screen for web based pilot study (3)

The participants were invited to answer memory and comprehension based questions on the text they had heard. The correct answers required a maximum of approximately 4 words to be

typed. The participants could also comment on the performance or the test by inserting a maximum of 255 characters including white space or leave it blank (Figure 37).

1. What was the name of the character speaking ?	Type the answer in the box.	<input type="text"/>
2. Who does she think she is talking to?	Type the answer in the box.	<input type="text"/>
3. What term did she choose to describe herself (her job)?	Type the answer in the box.	<input type="text"/>
4. In which episode did the heavy breathing occur?	Type the number in the box.	<input type="text"/>
5. What did 'he' say to her?	Type the answer in the box.	<input type="text"/>
6. What percentage of times is 'it' a complete disaster?	Type the number in the box.	<input type="text"/>
7. How long do you think the speech took from beginning to end?	Type the answer in the box.	<input type="text"/>
8. Did you think the character was reading or speaking ?	<input type="radio"/> Speaking	<input type="radio"/> Reading
9. Would you like to find out what happens to the character?	<input type="radio"/> Yes	<input type="radio"/> No
10. Please comment on the performance or the test	<input type="text"/>	
Thank you very much Please submit your answers	<input type="button" value="Submit"/>	

Figure 37: Questions screen for web based pilot study (3)

The unmodified synthesis of the text is (CD Track 65). The modified version is (CD Track 66).

char_name	who_talk	what_term	which_episode	he_say	complete_disast	beg_end	read_speak	happens_next	voice
JC	her agent	actroid	312	He loves her	61%	Five minutes	speak	N	kate
no idea	her agent	Actroid	312	That he loves her	61%	2-3 minutes	read	N	mary
don't know	answer machine	actor	312	Nothing	62%	5 minutes	read	N	kate
JC?	Her agent	Voice speaker	314	Nothing	62%	4:30mins	read	Y	mary
Jesse	Answerphone	Agent	312	Do i even know you?	66%	4 mins	speak	N	mary
Josaleine	Don't know	Actroid	312	I love you	99%	5 minutes	read	Y	mary
No idea	Her Agent	Acting Droid	312	get Lost	61	Too long	read	Y	kate
Don't know	No idea	Actress	312	That he loves her	72%	4 mins	speak	Y	kate
I Don't remember	a random person	call operator	215	he loves her	22%	10 mins	speak	N	mary
JC	her agent	Actroid	103	Don't remember	65ish	5 mins	speak	N	kate
JC 133	Don't know	Actroid	300+	If you ever call this#	61%	7 mins	read	Y	mary
Jane	A random person /staller	Actroid	363	Nothing	?	4 mins	speak	N	kate
Jodie	to the agent	actor	312	He loves me	61	5 minutes	read	N	mary
Jacile	Her agent	Actoid	312	I love you	66%	5 minutes	read	N	kate
Jose	You	? the operator	312	?	60%	5 min	read	Y	kate
Antoid? Android	her agent	an actoid	312	something about love	61%	5 minutes	read	Y	kate
?	Person with answers	Android unsure	312 or 322	?	67%	3 minutes	speak	Y	kate
Don't know	?	actoid	312	that he loved her	75%	3 minuits	read	N	kate
I don't know	Her agent	Actroid / actor	312	Don't know	42	4 mins	read	N	mary
Don't remember	a friend	actoid	312	he loves her	61	4 mins	read	N	kate
Mary?	Just some nameless customer	Actor	312	Don't call me again	61%	2 minutes	speak	Y	mary
Linda (?)	An answering machine	An Actroid	312	nothing	22%	1 min	read	Y	mary
Mary or Kate I guess	Agent	Actor	312	Don't remember	17	2 minutes	read	N	mary
Android	Her agent	Actor	312	I love you	61%	5 minutes	speak	Y	mary
	Boyfriend	I don't remember	312	I don't remember	61	20 minutes	read	N	kate
	Her agent Mts ?? (couldn't make out the r	Actroid	312	I don't remember and somethinh	66	4mins	speak	Y	mary
	Me	Lots, but she's an actor.	312	Don't know	Only a few	3 minutes	read	N	mary
	Her agent	Actroid	312	he loves her?	61	3 minutes	speak	N	kate
	Agent	Don't know	312	He loved her	61	2 minutes	speak	N	mary

Figure 38: Data collected from web-based pilot study (3)

Figure 38 presents a selection of the data collected, indicating the ranges of responses. Each column (1-9) represents the responses to one question. The field names at the head of the columns are contractions of questions 1-9 shown in Figure 37. Question 10 (used for participant comments) is not shown. The column on the far right tracks the modified (Kate⁶²) or unmodified voice (Mary). This information was not revealed to the participant. Table 29 shows an analysis of the data.

⁶² An arbitrary name: in retrospect it should have been PAT.

Question No	1	2	3	4	5	6	7	8	9
Correct keyword	JC	agent	actoid	312	He loves her	61%		Breakdown of speak/read	Breakdown of yes/no
Unmodified voice score (number of participants answering correctly)	3	6	7	11	7	8	72 minutes	Speak 6 Read 8	Yes 8 No 10
Modified voice score (number of participants answering correctly)	3	9	8	12	7	7	79.5 minutes	Speak 7 Read 11	Yes 9 No 8
P =	1.000	0.439	0.796	0.835	1.000	0.796			

Table 29: Results of web-based pilot study (3). The P row shows the p-value for a chi-square test of this data as reported by Excel software. Incorrect answers are not included in the calculation of the chi-square test.

For the analysis, use of the correct keywords was interpreted as a correct answer. Inconsistencies in spelling, and any extraneous words, were ignored.

The results (Table 29) show almost complete parity between the two versions of Microsoft Mary. In the comprehension test, the modified voice resulted in a total 46/108 (42%) correct answers, while the unmodified voice resulted in 42/102 (41%) correct answers; The p-value for a chi-square test of this data as reported by Excel software shows no statistical significance in these results. The results of the tests into user engagement (questions 7, 8 and 9) subject to the same statistical test, were also statistically insignificant.

Conclusion: Two conclusions may be drawn from this study. The first is that the reports of improved 'liveness' in previous studies by the application of the settings derived from the focus group and user survey are contradicted by this study. The second is that neither 'comprehension' nor 'engagement' (the dependent variables in this study) can be equated with 'liveness' and that the assumption stated outlining the goal of this study: *"something that had 'liveness' would be more engaging and therefore command more attention and be more memorable"* is incorrect. This outcome is related to a similar outcome from the first study in 'liveliness' 4.1, in which users identified more of a 'vague feeling' of a manifestation of 'liveliness' rather than a clear knowledge of what it was. This difficulty is addressed in subsequent framework studies in Chapter 7, in which a more thorough test of engagement is devised, and other indicators of the perception of 'liveness' are explored.

Before documenting the framework tests, the results of the tests so far lead to further consideration of each of the paralinguistic/prosodic modifiers implemented in the PAT framework, in studies in which user control over the settings is not provided and consequently predetermined settings must be established.

6.2 Setting the values for the prosodic modifiers

6.2.1 Empty pauses at natural punctuation points.

Of the useful studies, the focus group study, web-based pilot study (1) and the user survey all showed some evidence for user preference for empty pauses at natural punctuation points (e.g. full-stops, commas and question-marks). Question-marks were not specifically tested, but

there is no evidence known to the researcher to show that the duration of pauses at questions differ from the duration of other grammatical pauses, although common sense would suggest that a non-rhetorical question presented in conversation is likely to be followed by a pause long enough to allow for an answer. As the system is non-conversational, that possibility is ignored, and question marks and full stops are treated similarly.

The literature reviewed may be construed to confirm the results of the studies. An actor's dramatic pause will generally be situated at a full-stop. The Shakespearean verse-speaking method encourages changes of tone exclusively at full stops, with a significant change of tone generally preceded by a pause. Full stops may also be considered the orthographic equivalent of phrase boundaries in musical notation, which may not always be marked by a pause (unmeasured in musical terms) or a break (measured in musical terms), but either may be implied by the performer⁶³.

Evidence from linguistics suggests duration-ranges for empty grammatical pauses of between 250 and 1000 milliseconds (ms), with occasional longer pauses of up to 2000 ms, in order to approximate spontaneous speech. However, this does not take into account the performative 'over-acted' qualities inherent in the paradigm postulated herein in synthetic speech, and, accordingly, consistently-longer pauses, as well as very short pauses, were preferred in some studies (most notably, the focus group study).

The studies also confirmed the value of randomness in setting the duration of pauses. In all cases, predictable preset durations were rejected in favour of less deterministic settings. This aligns with the research reviewed in 2.5.8. into groove templates and humanisation algorithms, which exploit random perturbations to produce non-mechanical effects. There is some (weak) evidence from the web-based pilot study (3) that grammatical pauses may improve comprehension a little, but a better designed trial would be required to substantiate this. Additional evidence supporting this view is provided by (Whalen, Hoequist & Sheffert op. cit.) in a study on the effect of breath sounds in eliciting slight improvements in the retention of synthetic speech.

⁶³ Musical terminology tends toward the fuzzy and imprecise. Notation may be interpreted quite freely by performers; thus, a phrase boundary can be indicated by a break, pause, dynamic change, timbre change or any other method preferred by the performer and technically possible with the specific instrument employed.

Conclusion: The value for empty grammatical pauses should be set to a random duration between 250 and 2000 ms.

6.2.2 Empty non-grammatical pauses

The studies produced conflicting data in relation to empty non-grammatical pauses. Despite clear evidence in the literature that such pauses are commonplace in spontaneous human speech, attempting to reproduce this feature in synthesised speech appeared to have negative results. This is despite the fact that, in the focus group, the users had the option to iteratively grade the intrusiveness of the effect. The focus-group excluded empty non-grammatical pauses all together from their preferred algorithm, and users who took part in the survey selected the recording with non-grammatical pauses as the worst actor. Speaking speculatively, a reason for this could be that users accept hesitancy and inconsistency in human speech but find such features alarming (possibly indicating an error) in synthetic speech. Alternatively, the quality of rendering this feature using the PAT software tool produced an effect perceived more as glitch than a hesitation.

A more comprehensive study of this single feature might reveal convincing reasons to include empty non-grammatical pauses; however, due to their rejection by users in the studies described, non-grammatical pauses were excluded from the PAT framework tests.

Conclusion: Non-grammatical pauses should not be rendered in the formal PAT framework tests.

6.2.3 Breath filled pauses

Breath-filled pauses are commonplace in human speech, and improved rendering may yet make such pauses effective in synthetic speech. However, in these tests, and other trials in the framework tests (detailed in Chapter 7), this was not the case. Despite claims for better comprehension and recall by users when breath sounds are included in the speech signal (Whalen, Hoequist & Sheffert op. cit.), the subjects in the focus group unanimously rejected them and no further formal valid trails for their inclusion took place. Most users exposed to breath sounds expressed opinions correlating to the ‘Uncanny Valley’ problem. Alternatively,

the quality of rendering this feature using the PAT tool may have produced an undesirable effect.

Despite evidence in the literature for the paralinguistic expressiveness of breath sounds, users were discomforted by them within a synthetic speech stream. This may have been an outcome predicted by Mori's theory, but studies like that of (MacDorman 2006) continue to probe for the elusive 'edge' of the 'Uncanny Valley', indicating that the precise location of this 'edge' cannot be predicted reliably. Subsequent studies planned by the researcher (studies not documented in this thesis) consider the possibility of other sounds, more fitting for a 'machine', substituting for breaths, thus avoiding the 'Uncanny Valley' (See 9.6 Further research).

Conclusion: Breath-sounds should not be rendered in the formal PAT framework tests.

6.2.4 Periodic tempo (speech rate) variations

Periodic speech-rate variations were implemented in all the tests except web-based pilot study (1). Reasons for this omission are discussed in 6.1.7. The results appear to show a broadly positive response to the inclusion of speech-rate variations when rendered in combination with grammatical pauses (6.1.6 and 6.1.9), but not when presented in isolation (6.1.9). The metrics specified by the focus group and confirmed by the user survey are conservative (+ 7% fluency, - 5% hesitancy). It seems probable that the more extreme setting (+10% fluency, -10% hesitancy) designed to make the effect detectable was excessive when presented without the pauses framing the tempo changes in (6.1.9). This aligns with Renaissance verse-speaking features, like the results from the grammatical pause studies, to provide a 'change of tone', as specified by (Barton op. cit., Hall op. cit.).

Conclusion: periodic tempo variations were set as specified by the focus group: Fluency +7%. Hesitancy - 5%.

6.2.5 Background sounds

Of the studies documented in this chapter, only web-based pilot study (2) implemented background sounds. As the data from this experiment is unreliable, further examination of this

feature will be within the context of the PAT framework studies, the subject of the next chapter.

6.2.6 Choice of texts

The significance of the texts chosen in influencing the test results is a weakness in the experimental methodology that was assimilated retrospectively as the framework emerged. To ameliorate this problem, many of the tests compared one performance of a text to another performance of the same text. However, it cannot be claimed that the choice of voice and the choice of words did not undermine the reliability of the tests. As previously stated further work should be undertaken to consider the significance of the choice of texts, and work toward some sort of standardisation, or at least a categorisation.

6.3 Conclusions

In this chapter, a detailed account of the technical implementation of the paralinguistic/prosodic modifiers in the PAT software tool is provided. The metrics (including ranges, constants and the parameters for random permutations) were derived from a combination of the literature and user studies. Not all the user studies produced strong results, but of the five studies, three produced some consistent outcomes.

The final settings for the paralinguistic/prosodic modifiers for the PAT framework studies followed the focus group's recommendation. Web-based pilot study (1) and the user survey broadly confirmed the findings of the focus group study. The other two studies conducted produced unreliable data, but showed no underlying significant trend that should undermine confidence in the findings set out in this chapter.

The participants in three studies identified three features that improved their evaluation of liveness in a modified synthetic voice when compared to an unmodified synthetic voice:

1. Pauses at punctuated full stops and question marks of filtered random variable durations between 800 and 2300 ms⁶⁴.
2. Pauses at punctuated commas of filtered random variable durations between 375 and 600 ms (see footnote 64).
3. Consecutive periodic speech-rate variations at utterance-level between + 7% and -5% of Microsoft default speech rate of 4.21 syllables per second.

Although the results are interesting, the evaluations are not focused clearly on the concept of 'liveness' (the subject of this research). The participants reported satisfaction with the improvements they heard, but there is little evidence from these studies that what they heard was an improved manifestation of 'liveness'. Accordingly, an alternative set of studies are needed, framed in a context in which 'liveness' may be more apparent and measurable. The metrics specified in the results from the studies documented in this chapter provide the metrics for the implementation of the studies leading to the specification of the PAT framework: the subject of the next chapter. The PAT framework studies take the form of performances and installations.

⁶⁴ The reader is reminded that the Microsoft SAPI default value for pauses at full stops is appended by the PAT software tool.

7 Evaluating the PAT framework

“The most compelling interface s will make the user aware of her contexts and , in the process, redefine the contexts in which she and the interface together operate.”(Bolter & Gromala 2003, p.27)

The framework evolved as tests of manipulations made to the paralinguistic/prosodic modifiers in a synthetic speech stream using the PAT software tool indicated weaknesses in the initial assumptions and the potential significance of other variables on other dimensions became apparent. These alternative variables and dimensions are represented in the PAT framework. The evaluations are incorporated into an art exhibition, two telephonic art-and-performance works, and a theatrical performance. Methods are evolved to collect data assessing more complex levels of interactive engagement than those used in previous studies. The results reveal interesting but ambiguous trends in the user evaluation of synthetic speech artefacts.

7.1 Overview

In the last chapter, the studies leading to a set of metrics specifying the paralinguistic/prosodic modifiers in the PAT tool were documented. At a given juncture in the research it was hypothesised that the re-rendering of a synthetic speech stream, subject to appropriate user-defined settings and selection of the modifiers documented in Chapter 6, would be sufficient to manifest the perception of ‘liveness.’ The studies documented in chapter 6 did not show any evidence for a clear manifestation of ‘liveness’, and a gloss of the user comments reveal the possibility that the design of the experiments had constrained some user evaluations to a broad context of ‘naturalness’, or verisimilitude to human speech (see footnote 61), despite efforts to avoid this outcome. In order to provide a context more fitting to the evaluation of ‘liveness’, the PAT software tool was used to develop a number of alternative presentations in which the other framework dimensions would make a more significant contribution. Readers are reminded that two other dimensions - ‘place’ and ‘authenticity’ - are needed to complete

the framework for ‘liveness’ in synthetic speech. The temporal dimension is principally rendered by the modifications described in Chapter 6, but both ‘place’ and ‘authenticity’ require a broader approach more akin to a theatrical performance.

Thus, the emergence of the framework was a response to difficulties in realizing Proposition 1 (reproduced for convenience below) -

The first proposition is that the automated modification of pauses and speech rate variations can be shown to improve the user perception of liveness of a synthetic voice when compared to the same voice that has not been subject to such modifications.

- but also a return to first principles, and a reminder that the context for a synthetic voice is that of an actor with all the incumbent paraphernalia associated with performance.

Testing for the perception of ‘liveness’ in any context, including performance, is difficult. As previously discussed, the WSOD is negotiated by each user in the context of the specifics of the performance, and, accordingly, the perception of the degree of ‘liveness’ will vary between users. As ‘liveness’ is also an exotic term (and therefore not familiar to most users), asking a user if they experienced liveness, and how much, is not likely to produce any useful results. In order to arrive at a quantitative evaluation of the success of the framework, we needed to conduct a range of tests for different responses using different methodologies with the intention of triangulating the results. As previously stated this became an evolutionary process, using the outcomes from one test to inform the design of the next, with a consequent modification to the theory at each stage; a method with some similarities to ‘grounded theory’⁶⁵. The more pragmatic tests reviewed in Chapter 6 led to the more unusual qualitative tests reviewed in this chapter, and the progress was not as linear as implied by this documentation. At each stage of testing, the next stage was not apparent. For example, the first test, described in Chapter 4, was designed to determine whether users could detect ‘liveliness’ (not ‘liveness’) in speech and in speech-like tones (speech converted to melodic tones). The results suggested that this comparative test was limited, and the next study tested

⁶⁵ The method adopted in this research is not ‘grounded theory’. For more on ‘grounded theory’, see Cairns, P. & Cox, A. L. (2008). *Research methods for human-computer interaction*, Cambridge, UK; New York: Cambridge University Press. This research has made efforts to adopt the notion of iterative modification to the theory and the amelioration of the influence of the researcher in the process of qualitative research; however due to time constraints, the methods of coding and analysis found in ‘grounded theory’ were not considered for this research.

the ability of users to compose their own ‘lively’ speech sounds using a bespoke editing tool. This produced a positive affirmation for the prolongation and addition of pauses into the speech stream. However, one method of analysis of the results misdirected the research towards an algorithm for randomly-positioned pauses; a subsequent user survey showed these randomly-placed pauses to be ineffective. The concept of the PAT framework emerged when it became apparent that a number of related factors were contributing to the results for each test, and these factors could not be rendered using the voice-audio alone. Referring back to WSOD, it had been discovered that script and setting, as well as acting, were important contributors to WSOD, and therefore the perception of liveness was also subject to these factors. Further tests and performances documented in this chapter culminated in a theatrical performance in which the PAT software tool produced renderings of the Microsoft Mary voice which were presented to an audience alongside human performances. The audience was asked to evaluate the voices on a continuous ‘true’ or ‘false’ scale. The results of the tests showed that, although all the human voices were judged truer than the synthetic voices, differences in the perception of truthfulness of the same synthetic voice could be made if features mapped to the PAT dimensions were deliberately manipulated to propagate liveness.

7.1.1 The design of the PAT framework tests

In Chapter 2, we described a number of perspectives and sources informing the PAT framework but not directly contributing to the selection and setting of metrics; namely:

- Singing and opera
- Cinematic sounds
- Ventriloquism
- The ‘Uncanny Valley’
- Certain types of acting

These should be regarded as paradigmatic concepts that do not present clear-cut integration strategies or metrics within the context of a software tool or HCI framework. It is not possible to simply say: “Let’s make the voice more operatic,” or: “Let’s use background sounds as they do in the cinema”. Instead, it is important to identify the active operators or implications for user perception embedded in these concepts, and direct them at the problem of synthetic

speech. In other words, in order to account for the range of methods we used in the framework tests, we need to return to some of the useful perspectives from these sources and show how they could be exploited. In List 3 (reproduced for convenience below) it was suggested that the sources have the following perspectives in common:

1. The artefact does not disguise the application of a conceit, such that the user is fooled.
2. The user is expected to engage with an abstraction and to make the effort to flesh out the illusion for themselves using WSOD
3. The artefact may delight in demonstrating the mechanism behind the illusion when appropriate
4. The artefact may ascribe a source to itself. This does not have to be true.
5. The artefact may ascribe a location to itself. This does not have to be true.

These features form the basis for the design of multidimensional performance events; however, in accordance with the production of arts events, intuitive factors also played a part in the process.⁶⁶ Readers are also reminded that the PAT software tool and the PAT framework evolved together and not sequentially thus some features under development in the tool are trialed in the informal framework events although they subsequently dropped from the more formal tests. The first two events are reported in this document as ‘arts events’, as they would be in an arts publication. The audiences that attended these two events gathered as part of bigger related events designed for the dissemination of art and design. Accordingly there was no opportunity to conduct formal surveys or comprehensive user evaluations; it would have just been inappropriate. As an opportunity to informally test the software and hardware systems (subsequently employed in the more scientific studies) and to informally gauge responses they were very valuable. The other two events include scientific studies integrated into arts event. These studies are significantly more complex, and are reported in more detail, with some scientific scrutiny of the outcomes. However, it would be unrealistic to regard any of these studies as robustly scientific. They represent an alternative evaluation methodology where subjectivity and opinion is encouraged and the quest for hard data sublimated to a

⁶⁶ As stated in the preface, the convergence of arts practice with scientific research is deeply embedded in this research, but there is insufficient space to fully develop this theme. Of interest to the reader may be the numerous examples of collaborative arts science practice recorded in Wilson, S. (2002). *Information arts: intersection of art, science, and technology*, Cambridge, Mass.: MIT Press. and The Leonardo Journal Leonardo International Society for the Arts Sciences and Technology. Available at www.leonardo.info/ [Accessed 22/07/09].

holistic qualitative approach. One way to describe this would be that they help designers understand the problem-space through an informal discourse with users. As previously stated, by combining this approach with the approaches set out in Chapter 6, it may be possible to design synthesised speech artefacts within a multidisciplinary arts/science team that assimilate evaluation methodologies derived from both disciplines.

7.1.2 'Tide'

The 'Tide' installation was presented at the British HCI Conference, CCID symposium (Newell, Edwards, Andrews et al. 2006) as a collaborative artwork (See Figure 39). The goal was to explore the effect of clear but contradictory delineators of 'place', 'authenticity' and 'time' within the context of an art installation. Eight synthetic speech fragments were broadcast by eight pairs of paper-based artworks displayed on easels. The paper-based arts works juxtaposed a handmade quality with digital print-making techniques. The mode of broadcast was a miniature transducer attached to the paper, which produced a very low-level whispered effect that reverberated across the paper. Example recordings are included as (CD Track 73 to Track 77) however the acoustic properties of diffusion through the paper have been lost. The text used was the full length version of the text used in the focus group web-based pilot study (2) and the user survey. The text is included as Appendix J. The subject of the text was a man meditatively musing on stillness, silence and the dusk and the images reflected this to some degree. The prosodic modification was rendered according to the PAT tool settings with additional background effects, including sea-sounds and a clock. When appropriate, members of the audience were asked to comment on the artefact.

The specification for the eight pairs of recordings is listed in Table 30. All recordings had been prepared for Web-based pilot study (2) (See 6.1.8).

Recording	Recording specification	Notes	CD reference
1	Recording of synthetic voice; grammatical pauses (durations duplicated from the human recording)	Microsoft Sam	(CD Track 54)
2	Recording of synthetic voice; random-duration (between 500 and 2000 ms) pauses at full stops aligned to regular bar lines.	Microsoft Sam	(CD Track 55)
3	Recording of synthetic voice; random-duration (between 500 and 2000 ms) pauses at full stops; background sound (waves)	Microsoft Sam	(CD Track 57)
4	Recording of synthetic voice; random-duration (between 500 and 2000 ms) pauses at full stops	Microsoft Sam	(CD Track 58)
5	Recording of synthetic voice; random-duration (between 500 and 2000 ms) pauses at full stops; background sound (a clock ticking)	Microsoft Sam	(CD Track 59)
6	Recording of synthetic voice; no extra pauses	Microsoft Sam	(CD Track 60)
7	Recording of synthetic voice; random-duration (between 500 and 2000 ms) breath-filled pauses at full stops.	Microsoft Sam	(CD Track 61)
8	Recording of synthetic voice; random-duration (between 500 and 2000 ms) breath-filled pauses at full stops; background sounds (waves and clock)	Microsoft Sam	(CD Track 64)

Table 30: The recordings used in 'Tide'

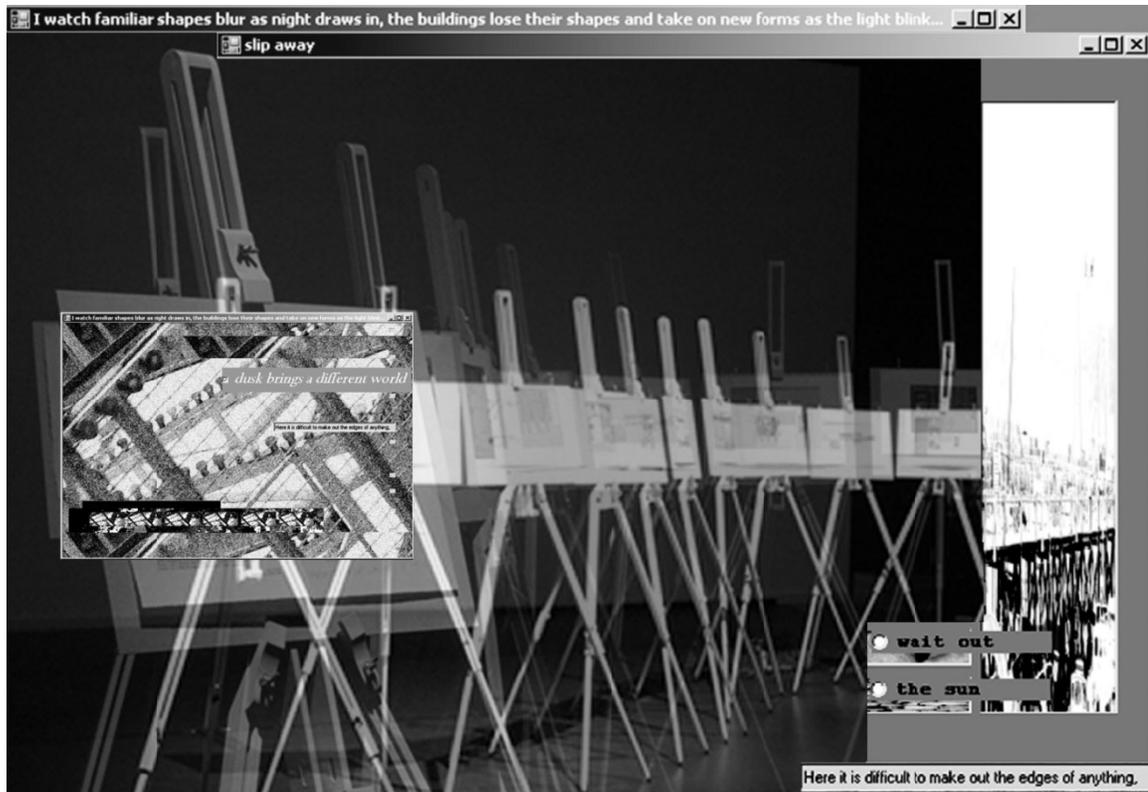


Figure 39: A composite image of the 'Tide' installation

This installation was set in a clear context of all three dimensions of the PAT framework.

1. 'Place' was defined (fictitiously) as either:
 - a. Inside the speaker's head as he thinks to himself.
 - b. By the sea.
 - c. Inside a room with a clock.
2. 'Place' was defined (literally) as an art exhibition.
3. 'Authenticity' was challenged by the juxtaposition of handmade and synthetic technologies as well as synthetic sound and real recorded sounds.
4. 'Time' was set by descriptions in the text which documented the passing of time, as well as by the metrical sound of clocks and waves, and by long sustained pauses rendered by the PAT software tool.



Figure 40: The 'Tide' artwork on display

An informal survey of the audience, based on individual conversations lasting between 30 seconds and 5 minutes, demonstrated the ease with which individuals could assimilate and accept the complex layers represented in the artefacts. Questions of naturalness, so prevalent in the studies presented in chapter 6 were not asked. The incongruous juxtaposition of real and synthetic sounds presented no difficulties, and the most used phrases were “haunting” or “atmospheric”.

Anecdotally, the ‘artistic’ context seemed to have alleviated anxiety and suspicion of synthetic speech, replacing it with broad acceptance and curiosity. By changing the context in which the voices are experienced by the user their expectations are changed. In this case an exhibition as ‘Place’ affords a different experience for the user, thus it argues (again, anecdotally) that an articulate PAT framework rendering, that takes account of place may support the user acceptance of synthetic speech. However, the highly specialised implementation does not support the notion of the general applicability of the framework.

7.1.3 'Call Centre'

The 'Call Centre' installation (Newell, Andrews, Edwards et al. 2007) was presented as part of a Digital Music Research Network conference, as a collaborative artwork. The objective was to gather user responses to breath-pauses (pauses with an audible breath sound) in a non-lab environment. Breath-pauses had been rejected by the focus group within a lab context. In addition, users would be exposed to a semi-comic or ironic text, based on a call center or telephone operator dialogues. The text is included as Appendix K. An old-fashioned phone was adapted to allow communication with the PAT software tool via MIDI (see 6.1.1); thus, the PAT tool rendered the paralinguistic/prosodic modifiers in near real-time according to user selections. The telephone would ring when a proximity detector picked up on a human presence, followed by a preset introductory speech. In this installation users could navigate the system themselves using the telephone keypad. The user could navigate the system freely or as instructed by the 'operator' voice. A typical sequence is included as (CD Track 78).

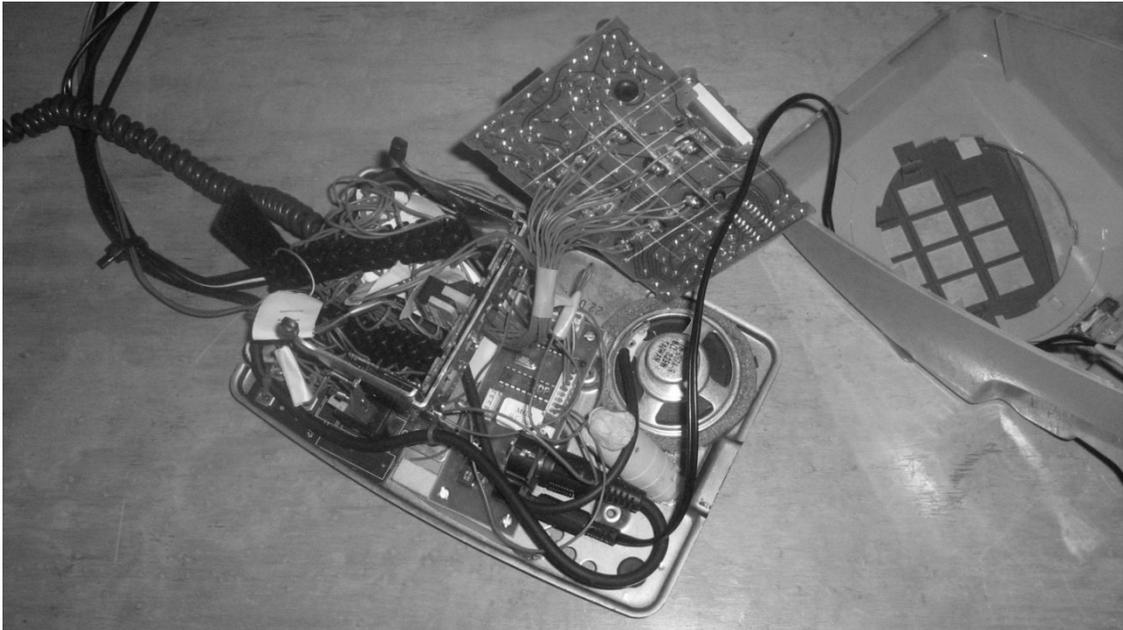


Figure 41: The adapted old-fashioned telephone used in 'Call Center.'

Again, this installation was set in a clear context of all three dimensions of the PAT framework.

1. 'Place' was defined (fictitiously) as both:
 - a. Inside a call centre.
 - b. On the end of a live telephone conversation.

2. 'Place' was defined (literally) as an art exhibition
3. 'Authenticity' was challenged by the juxtaposition of text materials, which displaced the authenticity of the 'place' (a possible thought-process: "it isn't really a call centre; therefore, the voices are inauthentic").
4. 'Time' was challenged by descriptions in the text documenting the passing of time, as well as pseudo-real-time responses, such as: "Where did you get to? I've been waiting hours! Hold on."

The rendering of the breath-sounds was set by the researcher. The relative volume, timing (pause-duration before and after the breath) and frequency were set intuitively to produce a nominally 'realistic effect' (i.e. the synthetic voice was breathing appropriately with the text). See section 6.1.2 for details of the breath rendering options in the PAT software tool.

An informal survey of the audience, based on individual conversations lasting between 30 seconds and 5 minutes, demonstrated the discomfort they felt with the breath sounds. There was no evidence that the breath sounds supported instances of 'liveness', rather the reverse applied with all those interviewed describing them as "eerie", "weird" or "scary". In other words, "uncanny". This confirmed the findings of the focus group, which had also rejected breath pauses.

7.1.4 'Please wait with me'

The 'Please wait with me' installation (Newell, Edwards & Andrews 2008) was presented at the University of Hull and at the SightSonic International Festival of Digital Arts, York, UK. This installation combined the features of an art installation with those of a scientific study, and, as such, will be documented in detail.



Figure 42: Post-card advertising the 'please wait with me' installation

Overview and goal: This art installation was designed to test the paralinguistic/prosodic modifier-settings derived from the previous studies in a full framework context; the voice would be mapped to the dimensions of 'place', 'authenticity' and 'time', and 'liveness' would be rendered through the script, the setting, and the acting. Although this was a full framework implementation, the framework was not being tested. Rather, the objective was to measure the user engagement with a synthetic speech artefact subject to the paralinguistic/prosodic modifier settings compared to the user engagement with the same artefact unmodified. This test was designed to resolve ambiguous results from the earlier studies (6.1.7 and 6.1.9), which demonstrated the positive perception of synthetic speech subject to the pauses inserted by the PAT software tool, and (6.1.10), which produced no useful data. The style of test was to be interactive, with users engaging with the voice as a piece of telephone-art. Crucially, the system would be tested in the field in an unmoderated environment, unattended and with no intervention possible by the researcher. The context for the user was similar to that of an interactive art exhibit or technology demonstration, as might be encountered at a museum or gallery. To facilitate this, a second old-fashioned telephone was adapted to allow communication with the PAT software tool, and additional features including speech

recognition and voice recording were added to enhance interactivity and data capture. A new script was commissioned. This is included as Appendix L. Engagement was measured in the following ways:

1. The total duration of the call, from picking the receiver up to putting it down.
2. The number of interactions with the system in the form of responses to synthetic voice instructions. These took the form of pressing a button and recording a message.
3. The duration of the interactions with the system in the form of the file sizes of the voice-messages left.

This method provided three engagement indicators of slightly different types. In this way, the expectation was that different personality types amongst the users would be accommodated within the test structure; impatient extroverts could leave a few long messages while more contemplative introverts might stay on-line a long time but leave short messages. The null hypothesis was that the modified speech would have no impact on these measures of user engagement.

Participants: A total of five hundred⁶⁷ individual pick-ups were recorded. The participants fell into two broad groups: three hundred and thirty six participants were located at the university campus, while one hundred and sixty four were visitors to a digital arts festival in a different city. In both cases, the artefact was presented in an environment in which the participant might 'stumble across it'. In other words, these were impromptu interactions between the user and the system. This may partly account for the brevity of some interactions. The two groups produced different results. No other information was formally gathered from users besides the three engagement-measures outlined above. Although anecdotal evidence suggests that some users used the system more than once, photographic evidence generated by an installed video camera in one location does not support this, and therefore it should not represent a serious spoiler. It is not possible to say how many of the first group had participated in other studies for this research. None from the second group had participated in the other studies.

Equipment: The interface to the PAT tool was a 1970-80s British push-button phone, chosen for its distinctive color and character in order to attract interest. Most of the internal

⁶⁷ Exactly 500 pickups for the two events together were recorded. Such a neat number was the result of chance.

components were removed and a MIDI interface converting resistance changes from the keypad to MIDI note-on/off data installed. In addition, the function button on the top of the phone and the receiver cradle were integrated into the MIDI interface circuitry. The MIDI interface board was adapted from the type customarily used to convert historic organs to MIDI instruments⁶⁸. By pressing any of the buttons on the phone, the user could trigger different responses from the system depending on the system state. The telephone was connected to a laptop running the PAT software tool. The tool was adapted to accept MIDI data to trigger the SSML processing and voice input to initialise the system and to initialise voice recordings. Thus the user was required to say something in order to start interacting with the system. An instruction to this effect was placed prominently on the telephone, although most users would say “Hello?” when answering the telephone. In the university environment, a video camera was installed that triggered a snapshot of each user engaging with the system in order to monitor repeat visits.

The user may interact with 1 or 2:

1. The Microsoft Mary voice, as specified in 6.1.3, with no modifications except extra two- and four-second pauses at strategic points in the dialogue allowing the user time to interact with the system by pressing the record button (these pauses are indicated by the SSML code included with the text as Appendix L). Depending on the speed of user interaction the duration of the sequence was approximately 15% shorter than item (2).
2. The Microsoft Mary voice processed with the paralinguistic/prosodic modifier settings defined at the conclusion of 6.2 plus additional two seconds and the four-second pauses at strategic points in the dialogue, allowing the user time to interact with the system by pressing the record button. Depending on the speed of user interaction the duration of the sequence was approximately 15% longer than item (1).

The settings were mock-ups of an old-fashioned interior with an easy-chair, rug and a few suitable props (see Figure 43). They were designed to both put the user at ease and to be sufficiently out of place with the environment to keep them aware that they were part of an art installation, with some expectation of participation. The two settings were different but designed to evoke the same atmosphere. In this respect, they were certainly not meant to

⁶⁸ The MIDI board was supplied by the MIDI Boutique in Bulgaria <http://www.midiboutique.com/>

represent a conventional setting for a usability test for interaction with a telephonic synthetic voice.

Procedure: The telephone would ring if pressure pads or a proximity detector detected a participant nearby. On picking up the receiver, the listener would hear the introduction to story of 'Catelina', the female voice on the end of the line. The full text is presented in Appendix L. To engage the listener, the system would offer a series of clearly bogus incentives to stay on-line, and play back recordings of other (fake) participants' supposed messages. To represent credible connection delays, 'musak' would occasionally be played. From time to time, the listener would be asked to leave a message or respond to a question. The progress of the story was from cool, professional objectivity to personal and emotional content. The longer the listener stayed on line, the deeper the story would go. Unlike 'Call Centre', all superfluous button press options resulted in an enquiry from the system: "You pressed button (x). Why?" prompting another message to be left. Messages could be left at any time, but users assumed they could only leave messages when instructed by the voice. The two versions of the speech alternated on pick-up; thus, a user would only experience either the PAT processed version or the original un-modified voice, unless they picked up again immediately after putting down the receiver. A recording of the entire sequence (the PAT modified version) is included as (CD Track 79).



Figure 43: The 'please wait with me' installation in use in a university corridor.



Figure 44: The 'please wait with me' installation at a digital arts festival

Results: Table 31 and Table 32 show the results of the 'please wait with me' installation at the university campus and at a digital arts festival.

Installation 1 (University Campus)

Voice	Total number of pick ups	Total call-duration = sum of all participant calls	Average call-duration	Total number of messages left	Average number of messages left	Total file size of engagements KB	Average file size of engagements KB
Microsoft Mary unmodified	171	02:18:15	00:00:49	151	0.88	5342.02	31.23
Microsoft Mary modified	165 ⁶⁹	1:27:12	00:00:32	82	0.49	3029.56	18.36
Microsoft Mary modified less 15% allowance		1:14:07	00:00:28				

Table 31: Results from the 'please wait with me' installation at the university campus

Installation 2 (Digital Arts Festival)

Voice	Total number of pick ups	Total call-duration = sum of all participant calls	Average call-duration	Total number of messages left	Average number of messages left	Total file size of engagements KB	Average file size of engagements KB
Microsoft Mary	82	1:11:48	00:01:08	37	0.45	754.7	9.20
Microsoft Mary modified	82	1:34:23	00:01:27	39	0.47	777.0	9.47
Microsoft Mary modified less 15% allowance		1:20:10	00:01:13				

Table 32: Results from the 'please wait with me' installation at a digital arts festival

The results show a reversal of expectation at the university campus, with evidence of less engagement by users in the modified voice than in the unmodified voice. All categories of

⁶⁹ The system recovered from errors six times. On recovery, it would reset to the unmodified voice.

engagement metrics reinforce the evidence for this trend. Although this result is reversed at the digital arts festival the evidence is weak. An allowance of 15% for the longer duration of the modified speech reduces the difference to a nominal amount as shown in the bottom row of Table 31 and Table 32. There are clear differences between the two events, in terms of user engagement, shown by the average call durations, however not between the two voices within the context of the same event.

Conclusion: The results from this study appear to identify problems in proving Proposition 1. It seems that the prosodic modifications that resulted in improved evaluations in the lab result in no improvements in the field and may have had the reverse effect. This result supports the weak results from (6.1.10), which also showed no improvement in engagement as a result of the PAT algorithm, although the study was flawed. This is further evidence that changes to the voice alone (to “how it speaks”) are not enough to improve user engagement with a synthetic speech artifact, and that the other dimensions described by the framework are likely to be as important. It may also indicate that, in interactive environments (as against passive listening environments), the insertion of extra pauses, hitherto considered a potentially positive enhancement, may be a distraction or an annoyance. This is an additional indicator that the context of use for the synthetic voice must be considered when designing appropriate paralinguistic/prosodic modifiers, and that any individual framework dimension or rendering method may not be effective in isolation. Testing full framework implementations and the framework itself is the subject of the next section.

7.1.5 ‘PAT Testing:’ testing the Framework

In ‘Please wait with me’, all the elements of the framework were presented, and one rendering technique - acting - on the ‘time’ dimension was tested. However, the framework itself was not tested.

The most obvious way of testing the framework would be to set a non-PAT framework synthetic voice against a PAT framework synthetic voice and ask users to evaluate the ‘liveness’ of each. To do so would require normative versions of a script, a setting and acting which could then be modified according to the framework principles for one voice and left as they are for the other voice. The problem with this approach is that, in practice, the normative assets

would have to be contrived to be less effective than the PAT versions in order to create sufficient distinctiveness for the user to evaluate. The test designers would in effect be creating a voice artefact that is specifically designed to fail. Thus the test would be self-fulfilling and very likely have a pre-determined outcome. Rather than adopting this approach, the 'PAT Testing' performances would be designed to demonstrate degrees of effectiveness of the framework as a direct result of design decisions made in accordance with the framework, rather than binary distinctiveness. In other words, all the examples tested are the product of framework manipulations (whether conscious or unconscious), and no specific effort is made to produce a normative one; but if variation in the user response can be predicted according to degrees of adherence to the framework principles then this may provide evidence of the framework's effectiveness. Thus, to show that the framework has value it needs to be shown that:

1. There is variation in the user perception of liveness for different voice artefacts.
2. That this variation can be measured.
3. That any positive variation is the result of the framework.

To facilitate this, a series of nine voice-artefacts were presented to a live theatre audience. These offered a range of different framework implementations. Some applied to human voices; others applied to synthetic voices. The live human voices were included to provide 'liveness' benchmarks, on the assumption that the human voice will always be perceived to exhibit more 'liveness' than a synthetic voice. One synthetic voice implementation was specifically designed to maximise 'liveness' using the framework (the design is detailed in 7.2), and to verify the predictive capabilities of the framework. The prediction was that this implementation would be awarded the highest 'liveness' rating of the synthetic voices, despite the voice being identical to two others being tested.



Figure 45: A postcard advertising the 'PAT Testing' performance.

Overview and goal: The 'PAT Testing' performance was intended as the culminating event of this research, in which a mixed context of scientific research and theatrical performance would be experienced by a live audience. The experimental materials were integrated within a performance that had to also serve as entertainment: thus a programme of performance pieces was arranged not all of which were strictly related to the prescriptive goals defined by this research. However embedded into the entertainment were three synthetic speech performances all of which were delivered by the same synthetic voice and thus any changes in the user evaluation would not be based on the voice per se, but on modifications to the framework dimensions to which the three performances were mapped. The goal was to establish the effectiveness of the PAT framework in changing the perception of 'liveness' in synthetic speech. The prediction was that the results would demonstrate complexity in the audience's comparative evaluation of human-versus-synthetic and synthetic-versus-synthetic speech when challenged to define a 'true' and a 'false' voice. This would substantiate the notion of relative 'liveness' posited by Auslander (op. cit.), and thus, equally, substantiate a

perception of 'liveness' for synthetic voices that could be subject to manipulation within the PAT framework. A broadly uniform response of computer-generated voices equating to 'false' and human-generated voice equating to 'true' forms the null hypothesis. An additional challenge was to see if any of the synthetic-speech-based works could achieve a higher liveness rating than any of the human-performed pieces, and, also, whether a specific voice artefact, contrived to optimise 'liveness', would achieve the highest rating. The prediction was that this would occur in the performance of 'Microsoft Mary's Comic Potential'.

Six separate performance pieces were presented. Some used human actors and voices; others used synthetic voices processed with the PAT software tool. The performance on DVD is included with this thesis. The performance ended with a short musical piece using materials presented earlier in the evening. This piece was not evaluated by the audience, and does not feature in this analysis.

This was an ambitious project, involving five actors, five computers, a series of stage settings, full stage lighting and sound. The mix of works and the voices were designed to be both entertaining and to expose the audience to a range of manifestations of voice-based 'liveness'. The continuum of voices ranged from:

1. Obviously synthesised speech
2. Disguised synthetic speech
3. Synthetic speech morphed with human speech
4. Amplified and processed human speech
5. Recordings of human speech
6. Unamplified human speech
7. Unscripted human speech

The independent variables operating within the performance were typical of any new media theatrical performance piece. No controls were exercised on potentially confounding variables. They included:

1. A range of different scripts in different styles
2. A range of different actors
3. Physically embodied voices (by actors)

4. Disembodied voices (on loudspeakers)
5. Computer-generated actors
6. Videos of actors

A troublesome issue was what question to ask the audience in order to generate a useful response. This issue was resolved in conversation with Paul Cairns (Cairns & Cox 2008) at the University of York and the decision was made to ask a very open-ended question: “Mark where you think the voice lies on the continuum from False to True.” This was agreed to be the nearest-achievable-measure of ‘liveness’ applicable to the wide range of experiences the audience were subject to in this study.

All the relevant support documentation is included in the appendix. The items are:

- An example report card (Appendix P).
- The slides presented to the audience to explain and instruct (Appendix T).
- The layout of the performance space and the equipment set-up (Appendix S).
- The scripts for each item are referenced, if a published work, or included in the appendix, if original work or fragments from a published work:
 - ‘Please wait with me’ (Appendix L)
 - ‘Not I’ (Beckett 1973)
 - ‘Microsoft Mary’s Comic Potential’ (Appendix N)
 - ‘Ohio Impromptu’ (Beckett 1990)
 - ‘A Duet: A Duologue’ fragment (Appendix M)
 - ‘Not I’ reprise fragment
- The programme notes (Appendix R).

Participants: Fifty four participants returned evaluation forms. Most of the participants were associated with the university in some way, either as members of staff or on the university performance mailing-list. Some may have participated in previous studies documented in this research, but none were members of the focus groups, or were known to have participated in the web-based pilot studies or the user questionnaire. None of the participants were direct contributors or performers in the production. All were above the age of 18.

Content: In the following list, where a performance has several voices they are separated by a '/' character; thus, the first performance has three voices: "1/1 human, 1/2 human recording and 1/3 synthetic".

Performance 1: 'Please wait with me' - The same material previously presented in the installation (see 7.1.4) was presented to the live audience. Two actors took the roles of the user listening and responding to the 'Catelina' voice, modified and rendered live by the PAT software tool. The audience could hear Catelina's and the actors' voices, as well as the recordings of fake users' messages. Thus, in this piece, three types of voice were heard: the actors improvising their responses in real-time; the recordings of fake users and Microsoft Mary playing the role of 'Catelina'. Accordingly, the audience was asked to evaluate three voices after this performance.

Voices: Perf 1/1 human, Perf 1/2 human recorded, Perf 1/3 synthetic.

Performance 2: 'Not I.' - Beckett's short play was performed in its entirety by a human actor. The play is written for a disembodied mouth, and only a mouth is seen on stage. The actor's voice was subtly amplified, just to make it more audible. The actor's eyes were obscured by sunglasses, providing a fairly convincing effect of a disembodied mouth. The actor was required to remain very still during the performance in order for nothing but the mouth to be fully lit throughout.

Voice: Perf 2/4 human amplified.

Performance 3: 'Microsoft Mary's Comic Potential' - This was performed by Microsoft Mary on a telephone answering machine. The speech was generated by the PAT software tool but not rendered in real-time, for technical reasons. Instead, the recording was captured on a hard disc player, and played back through a speaker located near the answer machine on the stage. This gave the effect that the voice was actually on the answer machine. The script is comic and the character is derived and adapted from the play *Comic Potential* (Ayckbourn 2002)

Voice: Perf 3/5 synthetic.

Performance 4: 'Ohio Impromptu' - Beckett's short play was performed in its entirety. In the play, a doppelganger of the single male actor is required to control the action by banging on

the table (he never speaks). This is strictly scripted, although the impression is quite random. In this version, the doppelganger was presented on a series of digitised film-loops that could be manually controlled to create a real-time effect.

Voice: Perf 4/6 human.

Performance 5: 'A Duet: a Duologue' - A fragment of the original play by Conan Doyle (Doyle 1903) was initially presented by a male and female actor, giving the impression that one or the other (or both) was about to speak the lines. A curtain was pulled in front of them before the voice was heard, thus the audience were unsure about the source of the sound. The sound broadcast was a mix of both human speech and synthetic speech, with the formants modified to produce a gender change using the TC-Helicon 'Voice Live' hardware (TC-Helicon op. cit.). The source voices were Cepstral Millie and Cepstral Lawrence (Cepstral 2006). Thus the male voice was a modified (gender changed) synthetic female morphed with a modified human female voice, and the female voice was a modified (gender changed) synthetic male morphed with a modified human male voice. The effect of this was of a synthetic human hybrid occupying some of the particularly unfamiliar in-between-ground identified by (MacDorman 2006)

Voices: Perf 5/7 hybrid morph high. Perf 5/8 hybrid morph low.

Performance 6: 'Not I' reprise - A fragment from the beginning of Beckett's play was performed by Microsoft Mary modified by the PAT tool. An animated lip-synched mouth replaced the human mouth described in the stage instructions. The voice was faded out after approximately 3 minutes. Figure 46 shows a frame of the animated mouth.

Voice: Perf 6/9 synthetic.



Figure 46: A still frame of the mouth used in 'Not I' reprise⁷⁰

Performance 7 - The musical piece was not subject to audience evaluation and is discounted in this research.

Procedure: This test took the form of a theatrical performance in a university studio performance space seating approximately 60 persons. The performance was at 7.30 in the evening. The audience was provided with a clip board and pen with the report cards attached. The program notes (Appendix R) were not distributed until after the performance, to discourage informed opinions. At designated points after each performance, audience lighting would be raised sufficient to allow the report card to be legible, and the audience was asked to evaluate whether a voice was true or false on a non-gradated 92 millimetre (mm)⁷¹ continuum (a pencil mark on a line). Instructions took the form of written instructions presented on a projection screen in the center of the stage, and a Microsoft Mary voice-over for screens 3 and 4. The screens are presented as Appendix T. The evaluations were sequential with no natural opportunity to modify the report card between events, although this was possible. The voices were presented in the range of different dramatic works specified above by professional and student performers. The material ranged from comic sketches to full, short dramatic works. A DVD recording of the performance is included as DVD 2 with this thesis.

⁷⁰ Image from a web source that subsequently became unavailable. (Based on the 'Preston Blair' phoneme series freely available in many formats on the worldwide web)

⁷¹ A photocopying error resulted in the intended 100 mm line being reduced in length. This had no impact on the results.

Results: 54 participants returned evaluation forms. The results are presented in Table 33 and Table 34 and Figure 47.

In Table 33 the results from the 'PAT Testing' performances are shown. Each voice is classified as human, synthetic or human/synthetic. Description, notes and confounding variables provide insight into the context for each of the performance evaluations. The sum of the measured marks on the continuum is provided and the truth-rating is calculated as a percentage. The rank as reported by the Excel software for each voice/performance (the lowest value designating the highest rank with ties given the same average rank). The Friedman test taken from the χ square distribution with 8 degrees of freedom showed at least one statistically significant variance by rank.

Table 34 shows graphs for the frequency distribution of the results for each individual voice. The x-axis shows the marked value measured on the continuum in mm divided in 10 mm units. The y-axis shows the number of participants making a mark within that unit. Thus in, 'human 1/1', the highest number of participants (approximately 16) placed their mark at around 70 mm.

Figure 47 shows the 'truth-rating' for each performance with the three identical synthetic voices in grey.

Perf/Voice	Voice type	Description	Confounding variables	Sum (mm)	Rank	Truth rating	Notes
1/1	Human	2 female actors improvising unamplified	First evaluation - order factor	3276	3.35	66.2%	Possibly some favour exercised from those acquainted with the students actors
1/2	Human recordings	Recordings of amateur voice actors reading from a script	Amateurish voices poor performances	2942	3.82	59.5%	Similar rating to event 4/6 may indicate an evaluation based more on content or source than on mediation
1/3	Synthetic	Microsoft Mary modified with PAT algorithm	Emotive text	1737	6.21	35.1%	Little difference between this event and event 5/7 which implies that human machine morphing makes little difference
2/4	Human	Female actor subtlety amplified	Challenging modernist text	2838	4.09	57.4%	Little difference between this and event 6 may indicate that subtle amplification does not make a difference
3/5	Synthetic	Microsoft Mary modified with PAT algorithm	Funny	2381	4.92	48.1%	The highest score of all the synthetic speech events may indicate that comedy makes a difference, or that the honesty implicit in the script and staging also helped
4/6	Human	Male Actor reading	Challenging text	2862	3.97	57.8%	
5/7	Human/synthetic	Morph high voice	Strange sound	1731	5.48	35%	See event 1/3 above
5/8	Human/synthetic	Morph low voice	Strange sound	2042	6.49	41.1%	
6/9	Synthetic	Microsoft Mary modified with PAT algorithm.	Artificial embodiment. Last evaluation - order factor	1498	6.67	30.3%	Despite the fact that this is the same voice as item 3/5 the crude embodiment, the content and possibly the position in the test sequence produced the worst score of all
Mean				2367		47.6%	Approximately equally true and false
The results of the Friedman Test as reported by the 'XLSTAT software' on the raw data showed; observed value 89.072, critical value 15.507, Degrees of freedom 8, p-value (two-tailed) <0.0001, alpha 0.05							

Table 33: Results of the 'PAT Testing' performances. The result of the Friedman Test as reported by the XLSTAT software is shown in the bottom row

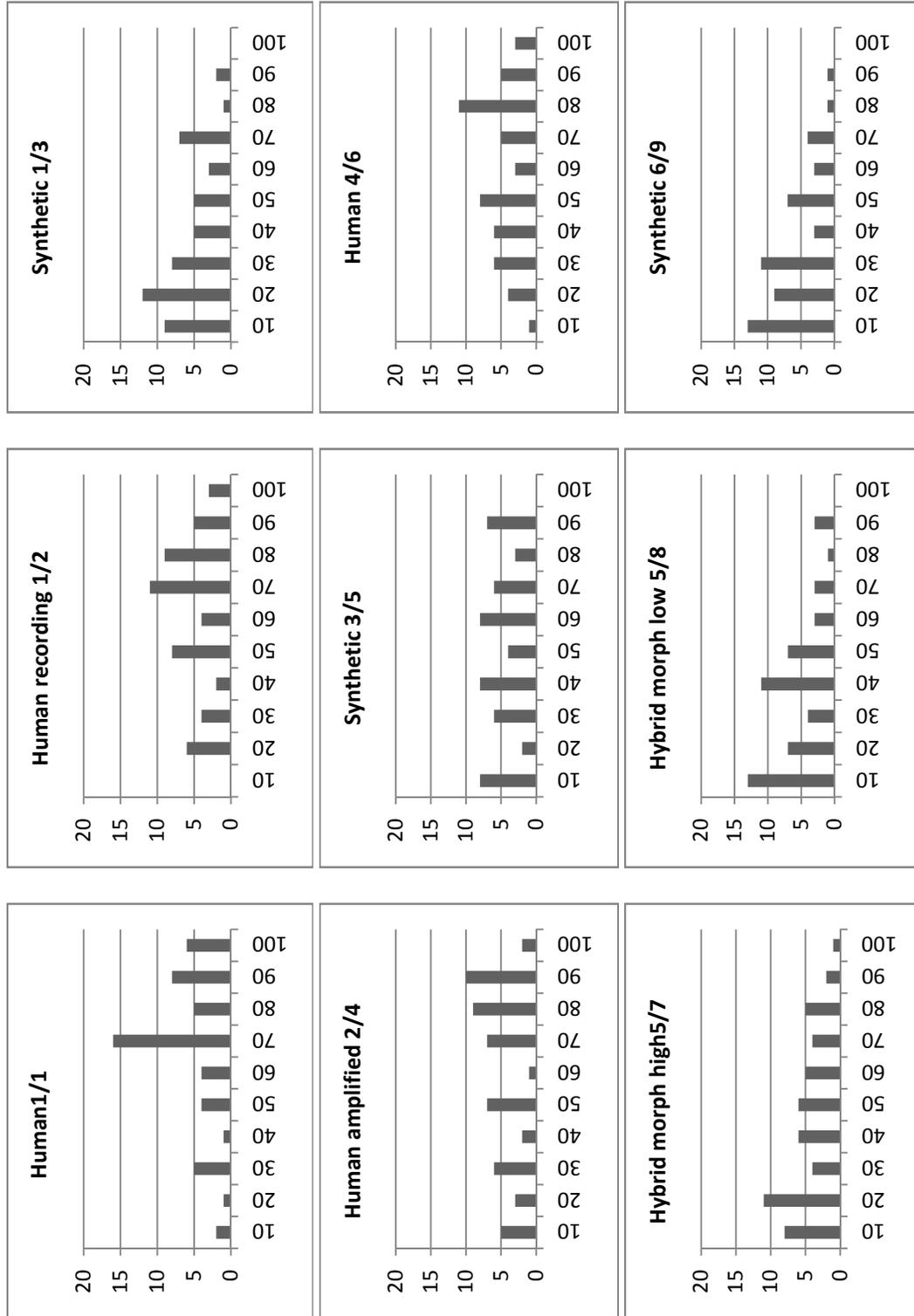


Table 34: Histograms of the results of the 'PAT Testing' performances. The x axis shows the score measured in mm: the y axis shows the number of participants

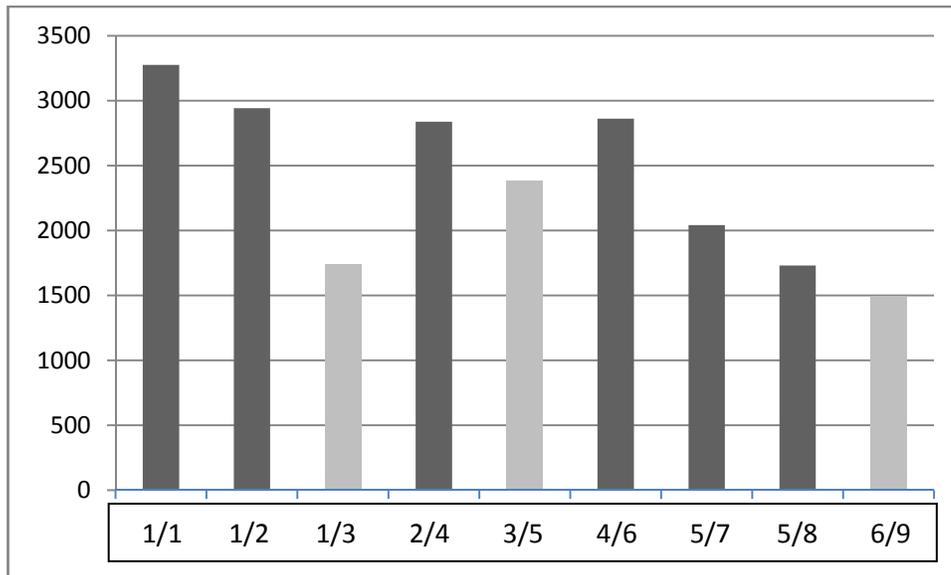


Figure 47: Chart showing the comparative truth rating of the 'PAT Testing' performances. The x axis shows the performance reference. The y axis shows the total score in mm.

As a 'truthfulness' benchmark, when challenged to associate a 'true' or a 'false' value in a performance environment, the participants awarded a maximum rating to a human performer of 66%. This contrasts with the lowest rating of 30%, awarded to the least true synthetic voice. Between these two extremes, the other human and synthetic voice truth-ratings are situated. The most successful of the synthetic voice performances was, as predicted 3/5, 'Microsoft Mary's Comic Potential' however the Dunn Multiple Comparison test as reported by Graph Pad Prism software showed a significance ($p < 0.05$) for 3/5 only when compared to 6/9 and not when compared to 1/3. As a number of particular confounding variables may have contributed to the low rating achieved by 6/9 (not least the order factor) this weakens the result. However the confounding variables in question are assimilated within the framework and their effect may be predicted as discussed in 7.2. The hybrid human/synthetic voices were not liked, showing a frequency distribution inclined to the left. This probably indicates an 'Uncanny Valley' issue. All the human voice artefacts have the slightly more normal frequency-distribution. The most successful synthetic voice artefact has a flat frequency-distribution, which indicates a wide range of opinions within a single performance and the possibility that the question was particularly difficult to answer in the context of this particular performance.

The order factor may be indicated by a tailing off of enthusiasm for any of the last three performances.

Conclusion: The most significant factor in this test was that 1/3, 3/5, 6/9 were the identical synthetic voice - Microsoft Mary modified with the same settings by the PAT software - but with significant optimisation of the framework co-ordinates applied to 3/5 (this is detailed in 7.2). Thus, it is shown that factors other than the voice-acting resulted in a difference of 18 percentage points in the truth-rating (the difference between the ratings for 3/5 and 6/9). None of the synthetic voice performances scored as highly as any human performance, but 3/5 - 'Microsoft Mary's Comic Potential' - was rated only 9.3 percentage points less true than its nearest human competitor. It may also be interesting to speculate on a 'truth ceiling' of 66% (the maximum achieved by a human performer) for any voice artefact in a performance context. This is despite the probability of a significant audience-bias towards the performers (as noted in Table 33), as their friends were present and may have felt an obligation to be positive. If 66% is the 'truth ceiling' then the achievement of performance 3/5 of 48% could be construed as a rating of 72% as true as the highest achievable truth rating.

Thus, in answer to the challenges posed at the beginning of this section, to show that the framework has value, it needs to be shown that:

1. There is variation in the user perception of liveness for different voice artefacts.

(Assuming the question presented to the users - "Is the voice true or false" - is a fair interpretation of the notion of liveness, variations in the user perception of 'liveness' have been shown in this test.)

2. That this variation can be measured.

(Assuming the method of measuring the user's evaluation - a pencil mark on a continuous line - is fair, the variation has been measured and some of the results have been shown to be statistically significant.)

3. That any positive variation is the result of the framework.

A prediction was made prior to the study that 3/5 would achieve the highest truth rating of the synthetic speech artefacts. Despite the range of confounding variables necessarily present in a

performance context, including the presence of human voice artefacts, this prediction was realised.

The 'PAT Testing' performance pitched human and synthetic voices together in a complex evaluative environment. The test demonstrated both the strengths and weaknesses of the PAT framework. Some robustness in the framework specification can be claimed, and the possibility of some predictive capabilities. The weaknesses exposed by the test lie in the responsibility left with the designer to specify the choices associated with each dimension and render method. This is an activity requiring significant creative input. Performance 3/5 'Microsoft Mary's Comic Potential', the most successful framework-performance, was the fruit of collaborative labour between a playwright of international renown, a professional actor and director. It is unlikely that many potential implementation environments will have similar resources.

One solution to the weakness set-out above would be if quantifiable values could be associated with the framework dimensions and render methods, providing a more robust predictive methodology.

In the next section, the issue of quantifiable predictions using the framework is addressed.

7.2 Can liveness be quantified?

This section is presented with the following caveat: in quantifying 'liveness' using the method outlined herein, a high degree of subjectivity could be hidden by the perceived rigour of the numerical representation. This is not the intention and accordingly no strong claims are made about the predictive capabilities of the framework.

The modifications to the framework coordinates can be independently scored using the following method.

A qualitative judgment is made for each of the render methods applied to each of the dimensions. The judgment is made according to principles set out in sections 7.1.1 and List 2 , and summarised for convenience below.

1. The artefact does not disguise the application of a conceit, such that the user is fooled.
2. The user is expected to engage with an abstraction and to make the effort to flesh out the illusion for themselves using WSOD
3. The artefact may delight in demonstrating the mechanism behind the illusion when appropriate.
4. The artefact may ascribe a source to itself. This does not have to be true.
5. The artefact may ascribe a location to itself. This does not have to be true.

A rudimentary scoring system would be:

- To award a score of 1 if the dimension and render method have successfully presented the principle (i.e. if the answer to the application of the principle is yes).
- To award a score of 0 if the dimension and render method have attempted to present the principle (i.e. if the answer to the application of the principle is uncertain).
- To award a score of - 1 if the dimension and render method have failed to present the principle (i.e. if the answer to the application of the principle is no).

The higher the total score, the better the framework principles have been applied, and the more 'liveness' the voice may present to the user. It is not always possible to provide an objective score for each render method and dimensions for each of the performances, and there is an element of arbitrary box-ticking that may confound the results, but, allowing for this caveat, it is possible to detect clear trends using this method. For example, a comparison of performance 3/5 and 6/9 reveals the following:

Performance 3/5 - 'Microsoft Mary's Comic Potential'

- In response to principle 1, 4 and 5 - She announces the conceit in the first few lines of text and nothing is disguised. The character of JC explains she is a robot (actoid), and this justifies the robotic voice style. She also explains where she is.
- In response to principle 2 - JC's comic timing is very poor; after all, her comic timing is derived aleatorically. The user must listen hard to catch the punch-lines, and this requires effort; however, this quirk also supports the humor making the overall effect funnier and demonstrates more liveness.

- In response to principle 3 - Some of JC's pauses serve the function of appearing to allow processing time for her to make calculations. Again, this is indicated in both the script and the acting, adding fictional credibility to her performance.

Performance 6/9 - 'Not I' reprise

- In response to principle 1, 4 and 5 – the character describes a fictitious place (probably an Irish rural community) discordant with the voice. Exercising WSOD in this circumstance is blocked by the dubious claims of authenticity in the script and the anachronistic voice.
- In response to principle 2 – The effort required to interpret the text, and the abstract literary code embedded in the text by the playwright, is excessive, and the reward insufficient to justify it.
- In response to principle 3 – There is no occasion when the mechanism is demonstrated and the audience let of the hook. Instead the conceit is relentlessly pursued, instilling weariness in the audience⁷². The animated mouth does not provide distraction; it merely augments the futile efforts applied to the conceit.

Table 35 and Table 36 illustrate the scores achieved by performances 3/5 and 6/9. Even allowing for the subjectivity applied to arriving at scores, a very clear prediction emerges. The prediction is confirmed by the empirical evidence in which performance 3/5 received the highest truth rating for synthetic voices and performance 6/9 the lowest.

	Place	Authenticity	Time		Sub Total
Setting	1	1	0		2
Scripting	1	1	0		2
Acting	1	1	1		3
Total					7

Table 35: Performance 3/5 scored for liveness

⁷² This effect was predicted; hence, performance 6/9 was curtailed in the performance after approximately 3 minutes. Of course this may either have confounded the result or it may have produced a less polarised result.

	Place	Authenticity	Time		Sub Total
Setting	-1	-1	-1		-3
Scripting	-1	0	-1		-2
Acting	-1	0	-1		-2
Total					-7

Table 36: Performance 6/9 scored for liveness

7.3 Conclusions

This chapter reviewed two exhibitions and two studies conducted to test and evaluate the paralinguistic/prosodic modifiers and the PAT framework. The framework test results provide some evidence that framework manipulations make a difference to the user perception of ‘liveness’; however, this must be seen in a context of some complex interrelations between the dimensions defined by the framework and confounding variables that naturally exist in the context of performance or exhibition. Although it may be fair to report that a performance or exhibition-space is not as controllable as a lab, it may be equally fair to report that the context of use for a framework for synthetic speech production finds a more appropriate analogue in a performance-space than a lab. Thus the results may be seen as more indicative of the type of results to be expected in a real-world application of the framework, and may have value as such.

In this research, the two exhibitions - ‘Tide’ and ‘Call Centre’ - were used to pilot the software and concepts in the field, before subjecting them to more rigorous tests. The exhibitions were also opportunities to develop the framework in the context of two real-world interdisciplinary collaborative designs. While it is not possible to make strong claims based on findings derived from these exhibitions, the informal contact they provided with users helped the researcher understand the wider reach of the problem-space. The installation ‘please wait with me’, and the performance ‘PAT Testing’, was an attempt to combine scientific rigour within the complex multidimensional interactive environment of live performance. Again, it would be wrong to

make strong claims, but some support for the thesis propositions have emerged from the studies documented in this chapter. This support is as follows:

- It is possible that optimisation of the single feature 'how a synthetic voice speaks' will not improve how a user evaluates it. This shown by the varied true/false evaluations for the same voice in 'PAT Testing'.
- The most significant user evaluation criterion of a synthetic speech artefact does not appear to be uniformly based on verisimilitude to human speech. This is shown by the poor true/false evaluations for the morphed human/synthetic voices in 'PAT Testing.' Arguably, the two voices were more human than the straight synthetic voices.
- Some of the techniques proposed within the context of the framework may improve user evaluations of synthetic voices. This was shown by the evidence that the voice designed to conform most comprehensively to the PAT framework was correctly predicted to achieve the best true/false evaluation.

A final outcome emerges that is not strictly relevant to the objectives of this research, but, for completeness, is worth including.

- Synthetic voices designed for interactive environments may have to meet different criteria from those designed for passive listening environments in order to be favourably evaluated. This is demonstrated by the failure to replicate the lab results showing improved user evaluations of a voice with extra pauses.

The proposition is that the PAT framework provides a conceptual model within which to negotiate the problem-space presented to designers of synthetic voices within an interdisciplinary design environment. Case studies supporting this proposition have been provided in this chapter.

In the next chapter, closer scrutiny of an imaginary design process is provided, in the form of hypothetical accounts of the design of three real-world applications using the framework.

8 Potential applications

In this chapter, three scenarios are envisaged in which an interdisciplinary design team uses the PAT framework to explore the problem-space for a specific TTS implementation. The first scenario is a telephone-banking-system. The second scenario is for a portable exhibit-guide-system for a technology museum. The third scenario is for a toy polar bear teaching aid. In all cases, the assumption is that the decision has been made to use TTS technology, rather than recorded human voices, at least for parts of the system implementation.

The following sections take the form of a series of hypothetical debates, as could take place at a meeting within an interdisciplinary design team preparing to design a synthetic speech artefact to a specific requirement. The general style is discursive, and is loosely based on some of the actual interdisciplinary design team meetings that contributed to this research, as documented by (Lopes 2009).

It may be helpful to the reader to imagine that, rather than pursue the (as noted) slippery objective of ‘liveness’, the fictitious design team tasked with these briefs have the set the primary objective as ‘truth’ as tested in the ‘PAT Testing’ performance evaluations in 7.1.5. Not all of the ideas presented in this discussion are presented as ‘good ideas’. In the imagined team-environment, the accumulated set of ideas is likely to be subject to some sort of selection process, probably in consultation with a client. One feature of the framework is to ‘frame’ the generation of design ideas so that the key components embedded in the problem space are revealed to the team.

In the PAT framework, the choice of synthetic speech technology, and thus the choice of voice to be subject to modifications using the framework dimensions, is largely irrelevant. In the design environment imagined in these exercises, that aspect of the design is assumed to be a given, in the same way that in the PAT framework tests the given voice was ‘Microsoft Mary’. Thus, in these case-studies, the reader should assume that the best technology, and therefore

best voice or voices has been selected, based on suitability for the design brief. The task for the design team is to optimise the truthfulness of the voice using the PAT framework.

As a reminder to readers, the approach to addressing the design problems set out in this section is derived from 'Technology as Experience' (McCarthy & Wright 2004, p.12), in which the authors claim that "...the quality of experience is as much about the imagination of the consumers as it is about the product they are using". The discursive, informal style is derived from (Sharp, Rogers & Preece op.cit. p. 48) Key terms, sources or perspectives discussed in previous chapters are referenced on the first occurrence in these scenarios.

8.1 Telephone banking

Brief: The brief for the telephone banking system is to design a TTS voice for users making transactions by telephone. The voice is not expected to handle general enquiries, but is required to negotiate customer transactions using (DTMF) touch-tone interaction. A typical, single-user, interaction may include keying in security information such as a pin code, transferring funds between accounts or making bill payments.

For this system, the team would be likely to favour a voice with intelligibility, clarity and precision over any other factors. However, all three dimensions of the PAT framework would need careful consideration by the team, even if the primary objective is to reinforce the first principles set out above.

Place: (Place, 2.5.9.2). A number of different fictitious 'places' are considered by the team. It is suggested that the 'place' defined by the voice could be busy and business-like to suggest a thriving business concern with a lively team of committed professionals. This could be done by using a number of different synthetic voices or genders to evoke a team, or by using background sounds evoking a busy office (Cinematic sounds, 2.5.10). One member suggests that this could be contrary to the principle of authenticity (Authenticity, 2.5.9.1), as it implies an imaginary human social environment. The team begins to favour a quiet, isolated environment in which security and privacy are paramount. In this case, the theory underpinning the rendering of the voice is aligned more closely to Chion's 'acousmètre' (Cinematic sounds 2.5.10), in that it operates displaced from the natural acoustic environment,

evoking power, authority and knowledge. Would it be possible for the user to choose or to be given different experiences, depending on where they are within the imagined auditory environment? For example, keying in a pin-code may be best conducted in a silent environment, while choosing between transfer options could take place in a soothing auditory environment. Someone makes the point that users are unlikely to be bothered enough to make a choice of auditory environment, given the goal of the interaction is to manage a bank account, but it is agreed that an automated process to evoke different places may help aurally render the conventional mental model⁷³ of a banking environment (likely to be applied by the user) and aid navigation.

The real place (location of the system) is the bank's I.T. centre, as opposed to the fictitious place (location implied by the system) which is to be the customer accounts department. Either of these two places could be rendered to optimise liveness. In this context, a distinctly non-human rendering of the prosody would probably evoke an I.T. centre and may be preferable to something less mechanical (more human) suggesting the customer accounts department. Affect (see structured vocabulary) is deemed inappropriate, in this context, and this further supports the notion of a standard voice (see structured vocabulary) with no aspirations to be affective, prompting a discussion on subtractive processes (see additive, distractive or subtractive processes 2.2) to rid the given voice of misleading emotional traits. Critically, the user's 'place' will normally be the relatively low-fidelity acoustic environment of the telephone receiver, located wherever they are making the call. A team member suggests that consideration of this factor is likely to be fundamental to the success of the voice design, as the environment in which this 'place' is experienced by the user will be subject to multiple unknown auditory interferences that would negate the effect of any complex and subtle renderings of place.

In accordance with this, the team suggests a quiet place with a soothing ambience. A voice in the female frequency range of the given synthesis system is chosen to overcome the poor rendering of low frequency sounds in telephone receivers. A metaphor modeled on a classical music radio station continuity studio is suggested.

⁷³ Mental models are an important concept in HCI. "Mental models have been postulated as internal constructions of some aspect of the external world that are manipulated, enabling predictions and inferences to be made." Sharp, H. Rogers, Y. & Preece, J. (2007). *Interaction design: beyond human-computer interaction*, Chichester: John Wiley. p.116 citing Craik.

Authenticity: (Authenticity, 2.5.9.1) The team agrees that ‘authenticity’ in the telephone bank environment is critical in maintaining trust and alleviating anxiety for the user. The ubiquitous cash machine has contributed to a culture in which interaction with a machine is quite normal for many users; thus, ‘authenticity’ may be maintained by a voice that is clearly machine-like and makes few concessions to naturalness (see structured vocabulary) or human verisimilitude. This supports the view put forward in the discussions on ‘place’. A long discussion on errors develops. Some members of the team suggest that trust⁷⁴ in the system is critical, and that any error should be immediately referred to a human advisor. Other members suggest that users are used to correcting errors made by machines, and that it would be annoying to be diverted out of a transaction just because the machine makes a minor error. ‘Authenticity’, in this context, is determined to be the speed and transparency with which an error is detected and referred back to the user for moderation. One team member suggests that another voice representing a supervisor could be used when an error is suspected. The team agrees, as this may disambiguate error messages from transaction messages, providing an effective user alert, and may also help retain ‘authenticity’.

Time: (Time, 2.5.9.3). The representation of real-time is also critical to the design of this voice. The team immediately agrees that an unambiguous notion of real-time should be expressed by the use of precise times, days and dates. In human speech, this would be excessive, but in a banking system it may be appropriate. One team member points out that this will have the effect of slowing ‘time’ down if the system is obliged to laboriously announce the long form of dates and times. A slower, relative ‘time’ is deemed acceptable for this system, but some team members are concerned that it could be annoying if the user has to make a number of transactions if on each occasion they have to endure the long form of the date and time. Yet again, the possibility of providing user options is discussed, and, yet again, this is declared the ‘last resort.’ Automated discretionary changes from the long to the shortened form are suggested, but this is declared to be in violation of the design principle of consistency⁷⁵. It is suggested that the solution in this case must rest with the script writer in the team, and that a

⁷⁴ ‘Trust’ is an important HCI concept. It is comprehensively addressed in Sears, A. & A.Jacko, J. (2008). *The human-computer interaction handbook: fundamentals, evolving technologies, and emerging applications*, New York: Lawrence Erlbaum Associates.

⁷⁵ Consistency is one of several design principles promoted by Donald Norman in his bestseller. Norman, D. A. (1998). *The design of everyday things*, London: MIT Press.

script be provided that strikes the right balance between unambiguous precision and the avoidance of tedium.

A team member points out that the representation of fictitious 'time' helps the user create a working mental model of the processes which conforms to inaccurate but useful (for the designer) preconceptions. For example a short delay of approximately two seconds after the voice has acknowledged a transfer arrangement submitted by the customer using DTMF would imply the execution of a process, and may accord reassurance to the user⁷⁶. The general speech rate and relative time should be slower than normal in order to meet the first principle - absolute clarity - and this may also create an aura of measured calm. Someone points out that the phrasing of numerical information has been shown to be critical in terms of retention⁷⁷ and that general assumptions that slow speech rate equate to intelligibility should not be made.

Despite this the team agree that measured calm neatly draws together a number of the themes arising during the discussion but add that any excessively long periods of silence should be covered with 'musak' or advertising. The notion of 'musak' also neatly captures the notion of the classical music radio station, previously considered an appropriate metaphor for the user experience. Beethoven's 'Moonlight Sonata' is suggested. It is finally agreed that some user choice could be appropriate in this regard, but the client is left with the choice whether to use advertising in order to re-inforce a potentially reassuring commercial framework (users expect respectable businesses to advertise) or the less invasive 'musak'. Table 37 shows the PAT framework applied to the telephone banking voice.

⁷⁶ This point is also made by Nass, C. & Brave, S. (2005). *Wired for speech: how voice activates and advances the human-computer relationship*, Cambridge, MA.: MIT Press. They suggest (P. 179) that "Even if the system could respond immediately, delaying for a second or two suggests that the system is taking the user's statement seriously."

⁷⁷ This point was made to the researcher by a delegate at Interspeech 2009, Brighton, UK who claimed to be an experienced speech-synthesis developer. It has not been substantiated and is thus reported anecdotally.

Dimension	Render method	Example outcome
Place	Script	'This voice is located in a secure server centre'
	Setting	An I.T. centre/a mobile phone handset/continuity studio for classical music station
	Acting	N/A
Authenticity	Script	'This is an automated speech service'
	Setting	DTMF pin code
	Acting	A female, mechanical voice quality Additional voice to handle errors
Time	Script	'Your transaction was executed at thirteen twenty-seven on July the sixth, two thousand and twelve'
	Setting	N/A
	Acting	'Your balance is <pause> thirty <pause> pounds.' Slower speech rate

Table 37: The PAT framework applied to a telephone banking voice

8.2 The technology museum

Brief: The system is to be installed on a hand-held device with a stereo earpiece for sound diffusion. The operating environment is a technology museum similar in content to the Science Museum in London⁷⁸, offering a range of contemporary and historic exhibits focused on science and technology. TTS has been chosen because the intention is to offer the user access to RSS streams and other dynamic networked resources, as well as the conventional scripted content. In any museum environment, the provision of a pleasing but informative experience is likely to be the first principle. While accuracy of information is important, a visitor using an aural guide will quickly abandon it for a conventional written guide if the experience is not pleasing. In a technology museum, the speech synthesis technology has the option to be an exhibit in its own right; in other words, it can be self referential.

The team discuss the issue of using pre-recorded samples for the non-dynamic content, and TTS for the dynamic. This seems to be the obvious way forward, but in the context of the demonstrative nature of the application it seems a shame not to use cutting-edge TTS

⁷⁸ More information on the Science Museum, London, UK may be found at <http://www.sciencemuseum.org.uk/>

technology to show off cutting-edge technologies. There are also issues of cost and ease of content updates to take into account if pre-recorded content is used.

Place: A number of options occur to the team in determining the optimal rendering of 'place'. One team member suggests that one solution is to imply that the voice is shadowing the visitor, and therefore always occupies the same aural 'place' as that of the visitor. Another team member points out that this can be facilitated by using binaural microphone technology to mix real-time ambience or to use pre-recorded ambience mixed into the speech stream. The rest of the team thinks this may be over-complicated and expensive. Another option would be to place the voice in the context of the exhibit's history, or in its own context. Thus, an exhibit showing a steam engine would merge the voice within a Victorian industrial sound-scape. Perhaps the safest option for 'place' would be, again, to position the voice, as Chion describes, as an instance of a type of 'acousmètre': in no specific 'place' but with omniscient (all-knowing) insights. A team member points out that a similar technique is used for voice-overs in television documentaries exploring an academic subject for a popular audience. The voice may be transposed from a university library to the fringes of the documentary world. The team find this option rather stuffy, and out of line with the more contemporary vision of technology the museum may wish to communicate. In the end, the team decides on a more radical proposal: a fictional place – the future – from where the voice is able to objectively review the achievements of past technologies as well as to speculate on future (from its perspective) innovations.

Authenticity: As one team member states, establishing 'authenticity' in the museum environment is likely to be informed by two factors: the authority embodied by the voice itself, and the appropriateness to the specifics of the museum experience. The team had already rejected the notion of an authoritative academic approach, despite the fact that it might have fitted some users' expectations of an electronic tour guide. In this context, a voice that is demonstrably trying to please and 'show off', as well as openly exposing the limits of its technological roots, may be fitting (see Brechtian 'alienation' (2.5.2) and Shakespeare's Theatre 2.5.3.) The fact that the voice's 'place' had been determined to be the future also provides more fanciful possibilities for features, such as characterizing the voice as a futuristic robot. It would be an opportunity for the technology to unashamedly over-extend itself, confident that the users are likely to understand why. Perversely, one team member comments, this is the

ideal opportunity to experiment with some of the more idiosyncratic anthropomorphic renderings (see extended vocal technique 2.5.6 and paralinguistic sounds 2.4.1), usually rejected on the grounds of absurdity. It seems more appropriate in this instance to provide greater user choice and options; thus further transforming the system from a tool into an exhibit or even a toy.

Time: For most users, a museum is expected to be a contemplative experience. A rushed school party would be an exception but they are less likely to make use of an audio guide. The speech should be designed to allow for contemplative silences, in which the user is encouraged to immerse themselves in the experience of the exhibit, and, in some cases, to interact with it. An interaction designer in the team pointed out that the speech should not be designed to be heard continuously, as it is likely that a user will frequently pause the system. This would create discontinuities in time that the system would either have to ignore or acknowledge. This could be done with the classic “Welcome back⁷⁹” gambit, but this can be clumsy. Time, and interruptions in time, could be accommodated conveniently within the notion of the voice that inhabits the future, and thus is in some way beyond time. Flexibility of time could include the past, as well as the present and the future, and could be evoked by changes in the style of language used in the script, or ‘graininess’, such as might be heard on an Edison wax cylinder or an early radio broadcast. Real-time could be user negotiated, by interrupting the playback, allowing for travel between exhibits. Alternatively, time between exhibits could be accompanied by synthetic voice chat or musings unconcerned with the museum context. This may evoke a pleasing browsing experience, as if with a friend, and eradicate the imperative to get to the next ‘important’ point or exhibit. The team agreed that, in this case, the voice artefact could be designed with a rich range of extra features targeted at a playful effect. No attempt at realism (see structured vocabulary) would be required, and this could be justified by making the artefact question its own authenticity, and by placing the voice in the future. Using this method, the system would be perceived as both a tool and an exhibit in its own right. Table 38 show the PAT framework applied to the museum guide voice.

⁷⁹ By “Welcome back gambit” the author is referring to the ubiquitous use of technologies on websites and computer games that allow the system to recognise a user and provide a customised welcome message.

Dimension	Render method	Example outcome
Place	Script	'My voice has been programmed by the future museum staff'
	Setting	A museum/headphones/the future
	Acting	N/A
Authenticity	Script	'You may adjust my speech rate if you wish'
	Setting	The future looking back
	Acting	A genderless voice with some unusual vocal characteristics
Time	Script	'This guided tour will be complete at about 12.35 your time'
	Setting	N/A
	Acting	Normal speech rate, long contemplative pauses Interruptible

Table 38: The PAT framework applied to a museum guide voce

8.3 A toy polar bear

This case-study is not realistic. It is very probable that a toy polar bear design team would implement a pre-recorded voice using an actor to impersonate a bear. It is included to demonstrate the value of the framework even in a domain where TTS would not be the technology of choice, as stated in 1.1.4.

The brief: The voice of the instructional toy polar bear has to be fun. As the target market is likely to be young children learning about wildlife or the North Polar Regions, it should also be informative. The anthropomorphic/ursine balance will be challenging to get right (the intelligibility of the voice versus the bear-like character voice (see structured vocabulary) it suggests).

Place: All the team agreed that this was an artefact with a number of options for the appropriate rendering of 'place'. The approach could be quite cinematic. The North Polar Region would be an obvious option, and provides plenty of scope for atmospheric weather sound effects such as rupturing ice flows, as well as paralinguistic extras, such as shivering sounds and teeth chattering. As the bear is primarily a toy, and the user is likely to be young, the effect of this must be tempered by fun and humour. One option would be to place the voice inside the bear, either in its vocal tract (it will be required to open its mouth to speak,

which could possibly be rather frightening) or, alternatively, in its stomach. This may concur with the expectations of some users, or at least those of their parents, gleaned from teddy bears who traditionally ‘vent,’ (the original meaning of which is “speak from the stomach”: see Ventriloquism 2.5.4). Someone suggests that perhaps the most potent render would be to place the voice in the user’s imagination. A secret conversation designed to exclude adults and non-bears. In this case, a close miked ‘intimate’ effect, or a sepulchral ‘lost in the mists of the imagination’ effect (see cinematic sounds 2.5.10), may be appropriate.

Authenticity: One team member stressed that ‘authenticity’ is no less a problem when the voice is a polar bear than a human. As previously stated, the user is likely to have strong preconceptions and expectations regarding an appropriate voice for a polar bear. These will be gleaned from a wide range of sources, from the zoo and animated characters, to images from books and from personal imagination. Once again, using Chion’s ‘acousmètre’ (see 2.5.10), it may be appropriate for the bear to know only what a bear would know. Of course, this is a fictional perspective. The bear may be understood by the user as a mechanical toy, in which case a mechanical twang may be acceptable. This may be an over-sophisticated expectation of small children, though. If the user has expectations of ‘bearishness’, the task of the script writer may be to persuade the user that ‘bearishness’ is not the same as a human impersonating a bear. If the script writer is successful, some of Wishart’s extended vocal technique (see 2.5.6), translated to a model of a bear-size vocal tract, or Shakespeare’s rhythmic versification (see 2.5.3), may not be out of place.

Time: This scenario lends itself to a flexible and playful approach to time that would be inappropriate for the bank’s telephone, and would require more effort to integrate into the technology museum guide. This polar bear character is positively enhanced by being make-believe, and this context frees the team to manipulate fictional time with little concern for real-time. Provided the user is informed of the dislocation of ‘real’ and ‘pretend’, time can become another toy. Thus ‘bedtime’, in the polar bear’s ‘time’, is not actually bedtime (in reality), but a shared temporal pretence between the child and the toy. ‘Time’ can also be disputed without detriment to ‘liveness’, because, in this case, the primary signifier of liveness is going to be in the mutual act of playfulness; and playfulness has rules agreed upon in a context of WSOD. Table 39 shows the PAT framework applied to a toy polar bear voice.

Dimension	Render method	Example outcome
Place	Script	'Hi, let's pretend it's really cold'
	Setting	The vocal tract of a teddy bear/loudspeaker
	Acting	Brrr... (sounds to indicate that it is cold)
Authenticity	Script	'Sometimes I say things in a silly way. That because I really speak "Bearish"'
	Setting	Styled as a toy polar bear
	Acting	A gender selectable voice with growl effects and resonance applied
Time	Script	'Today is my favourite day of the week. What's yours?'
	Setting	Tired voice at bedtime
	Acting	Frequent periodic speech rate variations and dynamic contrasts

Table 39: The PAT framework applied to the voice of a toy polar bear

8.4 Conclusions

The examples in this chapter offer a speculative snapshot of the framework in use. The analysis of the three dimensional rendering for each implementation is highly subjective, and there is no certainty that solutions similar to the ones proposed would occur in the real world. The significance lies in the encouragement to consider the three dimensions, and the choice of content for the three render methods applicable to each dimension. It is this process, and the manifestation of options (hopefully many and varied), that should lead to better design solutions.

The reader may note that the solutions are unusual, and this is an additional prediction for the framework: that it will lead to innovative solutions, with all the inherent risk associated with innovation.

9 Conclusion and future research

*“No attempt at natural, everyday speech should be made”
Constant Coquelin(1887) in ‘The Dual Personality of the Actor.(Benedetti
p.94 op. cit.)*

*“Exactitude is not truth” : Sometimes a less accurate mirroring of the world
can in fact be more effective.(Nass & Brave 2005, p.144. Citing Henri
Matisse)*

Chapter 9 draws together the results of the research, relates them to the three propositions set out in the introduction and provides the final conclusions. A short section considering future research concludes the thesis.

9.1 Conclusions to the three propositions

The three propositions set-out in 1.1.10 are reproduced for convenience below.

- **The first proposition is that the automated modification of pauses and speech rate variations can be shown to improve the user perception of liveness of a synthetic voice when compared to the same voice that has not been subject to such modifications.**
- **The second proposition is that the PAT framework provides a conceptual model within which to negotiate the problem space presented to designers of synthetic voices in an interdisciplinary environment.**
- **The third proposition is that the PAT framework can be shown to improve the user perception of liveness in a synthetic voice artefact when compared to the same voice that has not been subject to the modifications rendered by the PAT framework.**

9.2 Conclusions to Proposition 1

The first proposition is that the automated modification of pauses and speech rate variations can be shown to improve the user perception of liveness of a synthetic voice when compared to the same voice that has not been subject to such modifications.

A discussion on the suitability of the term ‘liveness’ in this proposition is documented in 9.5.1.

In Chapter 5, the sources and perspectives derived from Chapter 2 were tested and developed into metrics for the paralinguistic/prosodic modifiers implemented in the PAT software tool to produce automated modification of pauses and speech rate variations. A range of studies on the low-level parameter settings were conducted before embarking on the more comprehensive framework tests documented in Chapter 7. The results from tests on the PAT software tool demonstrated some improvement in the user evaluation of a synthetic voice on a comparative basis (one voice modified using the paralinguistic/prosodic modifiers compared against another unmodified voice). In particular, in the user survey (see 6.1.9), users declared that the Microsoft Mary voice, subject to the PAT modifications, was more than twice as good an actor as the unmodified voice. Despite this positive outcome, when scrutinised for more tangible indicators of appeal that would provide additional evidence of ‘liveness’ (an increase in the listening time demonstrated by users and interactive engagement activities such as button presses or leaving speech messages as instructed by the system) documented in 7.1.4, this result could not be reproduced. Outside of the controlled experimental environment, the effect of the PAT software tool automated modifications showed no appreciable increase in the perception of ‘liveness’ between a modified version of Microsoft Mary and an unmodified version. The application of pauses and speech-rate variations do not appear to produce a general improvement to the ‘liveness’ of the Microsoft Mary synthetic voice, except in a lab based comparative test where confounding variables may be controlled.

The studies conducted so far show variation in the user response to a synthetic voice, based on issues that can be defined only within the context of the full framework implementation. Thus, by extension, it may be concluded that without regard to the multidimensional rendering required for any performance by a speech artefact implied by the framework, modifications to

pauses and speech rate variations will not reliably improve the liveness of a synthetic voice. However, they may contribute to improved liveness when applied within this wider context.

Proposition 1 is true, within a strictly comparative domain (one voice, PAT-modified, judged against another, unmodified, voice) in a laboratory controlled test, but false in a broader implementation context, in the field.

9.3 Conclusions to Proposition 2

The second proposition is that the PAT framework provides a conceptual model within which to negotiate the problem space presented to designers of synthetic voices in an interdisciplinary environment.

In Chapter 7, the PAT framework was implemented, and in Chapter 8 speculative case studies were set out. In both cases the PAT framework conceptual model was applied and in 7.1.5 tangible effects of the framework modifications were recorded. In each of the implementations reviewed in Chapter 7, the conceptual model presented by the framework was applied by a multidisciplinary team engaged in a voice artefact design process⁸⁰. The artefacts constructed were tested in a range of environments, and some user evaluations were recorded, formally and informally. There is some evidence of successful implementations of the framework in 7.1.5 (Microsoft Mary's Comic Potential), from which it may be inferred that the positive user response was as a direct result of an appropriate application of the conceptual model embedded in the framework. Theoretical metrics demonstrated in 7.2 suggest that a qualitative analysis of the artefact, based on evidence of rendering of each of the PAT dimensions, may indicate some predictive capabilities in the framework, but this is not a strong claim.

Proposition 2 is true. The PAT framework provides a conceptual model within which to negotiate the problem-space presented to designers of synthetic voices in an interdisciplinary

⁸⁰ The design process leading to the 'Tide' project (see 7.1.2) was documented as the primary case study in Lopes, A. (2009). Design as Dialogue: Encouraging and Facilitating Interdisciplinary Collaboration. *Design Principles and Practices*, 3, 261-276. Lopes's findings and concerns are not relevant to this research but the reference is included for completeness.

environment. The usefulness of the model in the broader professional context is, of course, another question and cannot be demonstrated at present.

9.4 Conclusions to Proposition 3

The third proposition is that the PAT framework can be shown to improve the 'liveness' of a synthetic voice when compared to the same voice that has not been subject to the modifications rendered by the PAT framework.

The PAT framework tests applied the concept of the relative truthfulness of a synthetic voice to the problem of the evaluation of 'liveness' set out in this proposition. In the PAT Tests (see 7.1.5), it was shown that the user evaluation of the *same* synthetic voice on a true-false scale could fluctuate by 17% when subject to modification of the framework dimensions. A rigorous alignment to the PAT framework dimensions produced the most successful outcome for the synthetic voice tests, in the form of 'Microsoft Mary's Comic Potential'. The PAT framework is necessarily subject to many variables; thus, strong claims for the effectiveness of the manipulation of any particular variables on any particular dimension cannot be made. The contribution offered herein is to define a framework in which a small set of categories of variables are presented and a possible method of rendering each is suggested. The designer is still required to make critical judgments that are likely to have an effect on the success of the final design.

Thus, within this constrained context and subject to the specific design judgments made for these studies it may be claimed that ***Proposition 3 is true, and the framework can be shown to improve the liveness of a synthetic voice when compared to the same voice that has not been subject to the modifications rendered by the PAT framework.***

9.5 General conclusions

Designing synthetic voices that people want to listen to and potentially interact with is a difficult task. For reasons already stated, the acceptance of a synthetic voice is severely limited by the inherent sensitivity the human ear has acquired to falseness and artificiality in the

speech signal. It is not a unique problem, but belongs to a set of problems which coalesce around particular sensitivities that are powerfully encoded in human evolutionary development. These sensitivities, which incorporate all the senses, are critical to our capacity to avoid deception, and cannot be easily 'switched off'. This research has shown that there are at least two ways to address this problem for synthetic speech. The first is to continue to research ways in which the synthetic speech signal can more closely match that of a human. This remains the current direction of choice within the research community, and there have been, and are likely to continue to be, many advances made. The second way (explored herein) is to try to find other ways of making the inevitable falseness exposed by the 'Uncanny Valley' problem more acceptable, by focusing on 'liveness' rather than verisimilitude to human speech. The PAT framework helps the designer negotiate the choices presented within the complex problem-space presented by the second alternative; choices such as: "How much to pretend?" (expressed by 'authenticity'); "Where does the user think the voice is coming from?" (expressed by 'place'); and "Is this voice speaking now or is it a recording?" (expressed by 'time').

The context in which the second alternative has been explored is interdisciplinary and has been drawn from the performing arts, where many traditions designed to disguise falseness have evolved. The researcher believes that many of these traditions are so obvious that it has been easy to overlook their potential usefulness within this domain. Performance has always been required to plunder pretence, illusion, story-telling, abstraction and all the rich potential of the human imagination in order to be effective. There seems to be no reason why synthetic speech research should not also make use of all available resources, including manipulating the 'willing suspension of disbelief', to bring about synthetic speech that has more 'liveness'.

By applying a discriminating scientific approach, and seeking to focus on the few manageable and quantifiable variables revealed in the literature (principally periods of silence, or pauses), it was hoped that the sort of strong results expected in a computer science thesis could be reported. In actuality, most of the empirical results are not strong, and the findings proved more likely to confirm the complexity and ambiguities inherent in the problem of synthetic speech than to provide answers.

One strong result appears to be that the evaluation of an individual synthetic voice can be confounded on a true-false rating by other factors other than verisimilitude to human speech.

Many of these factors appear to lie in the specific context of use, and are therefore difficult for the designer to exploit. Factors like what the voice is required to say, what the setting of the voice may be, and, of course, who it's talking to, may significantly modify the user's evaluation of, and subsequent belief in, the voice. Unlike a computer-based system, the human speech system is able to adapt to these context-specific variables at multiple levels. Adjustment may occur at the linguistic level by modifying the choice of words, by choosing words appropriate for the context, or by modifying the tone of voice. Paralinguistic variables may be exploited, such as a well-timed laugh or hesitation. Computing these factors is significantly beyond the capabilities of current computer-based intelligence, and is likely to be so for some time to come.

This requirement for ambiguous and unpredictable adaptation is very troublesome for computer science because rules cannot be developed along the lines of 'if you want this kind of synthetic voice then do this; otherwise, do that'. The designer is prevented from closely defining the 'this' and the 'that' by a daunting set of choices that seem difficult to comprehend. The PAT framework offers a way of comprehending the choices by transforming a problem-space conventionally understood as scientific or technical into a problem space that benefits from a multidisciplinary team approach. By restructuring the approach to the problem, solutions may appear that at first-sight seem counter-intuitive but on closer examination set more pragmatic goals and make more economical (and effective) use of available resources, such as WSOD, than the prevalent anthropocentric approach.

9.5.1 Addressing the title of the thesis

The deconstruction of 'liveness', and the application of a scientific granularity to the concept such that a strict scheme of methods can be applied to an artefact in order to engender 'liveness', has been developed in the PAT framework. By implication, the title of this research posits 'liveness' as a desirable feature to incorporate into synthetic voices. By so doing, it attempts to circumvent the problems of 'realism' so vividly captured in the notion of the 'Uncanny Valley'. 'Liveness' is a slippery and elusive concept. It is arguable whether it is the best descriptor of the auditory features this thesis has striven to manifest in synthetic speech. In the studies, it has been subject to numerous re-wordings and adaptations in order to make it detectable and reportable by users in a variety of evaluation contexts, namely:

- ‘Lively’ (see 4.1)
- ‘Live’ (see 6.1.7)
- Sounding ‘spontaneous’ (see 6.1.8, 6.1.10, 6.1.9)
- The difference between ‘reading’ and ‘speaking’ (see 6.1.8, 6.1.10, 6.1.9)
- The ‘best actor’ (see 6.1.8, 6.1.10, 6.1.9)

In the end, the bare, crude terms ‘true’ and ‘false’ (see 7.1.5) may have been the most effective at communicating ‘liveness’ to the participants; however, to have focused from the outset on designing ‘true’ and avoiding ‘false’ synthetic voices would probably have led, at least in the minds of users, to confusion with ethical concerns. Conversely, the value of the term ‘liveness’ may be in its elusiveness, causing the synthetic voice designer, as a crucial part of the design process, to consider the slippery and unreliable responses exposure to a synthetic voice may prompt from users once the design has left the lab.

Thus, the use of ‘liveness’ in the title ‘Place, Authenticity and Time: a framework for liveness in synthetic speech’ could be considered a ‘stimulus’, encouraging designers to think about the design of a speech artefact as an elusive and problematic instance of performance-practice, as well as a technical and scientific challenge. The word ‘framework’ may offer solace from the otherwise hostile ambiguity implied by the use of a largely unfamiliar term from performance literature in a computer science thesis. The descriptors ‘Place’, ‘Authenticity’ and ‘Time’ spell out the intention to proffer a structured solution, but, most significantly, the title is intended to direct the focus of synthetic speech design away from reductionism, towards a more holistic approach that accepts ambiguities as an essential part of the design process.

9.5.2 The contribution of this research

This research has made a number of modest contributions. These are set out in three categories. The first set relate to new perspectives on the problem (numbered (I) – (II)), while the second set relate to specific outcomes from the studies (numbered (1) – (3)). The third set are presented as lightweight suggestions for which, at present, insufficient evidence has been gathered (numbered (a) – (d)). These may stimulate new design experiments and research proposals.

9.5.2.1 *New perspectives*

(I) **Interdisciplinary research environment:** The hope of the researcher is that the framework will provide a starting point for developments in a new field of interdisciplinary research in synthetic speech design and development (see Further research 9.6).

(II) **Less sophisticated solutions:** The failure to point the way towards a simple set of guidelines to improve ‘how a computer speaks’ by modifying the prosody may, perversely, be the significant contribution made by this research. If it proves to be generalisable (only time will tell), and, consequently, it is shown that the acceptability of synthetic speech cannot be significantly improved by ‘improving’ the voice (making it more human-like), then researchers may do best to wait until models for computer intelligence have progressed before hoping for progress towards better models of synthetic speech. In the meantime, there is all the more reason to look at alternatives that are not tied to advances at the highest levels of aspirational computer science.

9.5.2.2 *Contributions derived from the PAT framework*

Beyond proving the difficulties inherent in a strict technological solution, and pointing the way toward an interdisciplinary framework for synthetic voice design, the main contributions of the framework are as follows:

- (1) The PAT framework’s key contribution is to provide the synthetic speech developer with a means of mapping the user’s auditory interaction with an artificial voice on three continua: ‘place’, ‘authenticity’ and ‘time’. These criteria are distinct from conventional criteria such as naturalness or realism. Although precise positioning of the voice on the continua represented by these dimensions is not possible, an approximate positioning does prove to have an effect, and in 7.2, the effect may be shown to be (theoretically) predictable.
- (2) The second contribution relates to a need for a holistic approach to the evaluation of the user acceptance of a synthetic voice. In particular, insights gleaned from the tests into the user evaluation of speech artefacts as ‘true’ or ‘false’ in ‘PAT Testing’. The researcher agrees with Nass & Brave (Nass & Brave op. cit.) that because any speech artefact is necessarily and unavoidably representing a human, the user response is similar to the response to another human, although more extreme. This was revealed in the PAT tests which tested human voices alongside synthetic voices and showed a wider spread of results (18 percentage points) for the synthetic voices (see 7.1.5) than the human voices (9

percentage points). In accordance with this, the ambiguous results from the PAT tests (7.1.5) on the Microsoft Mary voice reveal the potential for the user to evaluate the artefact just as they would a person:

- On emotive, gut-instinct reasons, which are difficult to objectively rationalise.
- On what it says.
- On how it says it.
- On the place (or context) it says it.
- On a combination of the above factors.

To maximise the usefulness of the evaluation of a synthetic speech artefact, design and evaluative processes should be developed that allow for the existence of confounding variables such as these. One way of doing this is to use the techniques derived from theatre and performance documented herein and pioneered by Alan F. Newell.

- (3) The final contribution derived from the framework is the development of a preliminary structured vocabulary for the evaluation of synthetic speech artefacts. This could form the basis for further interdisciplinary discussion from which a more comprehensive model may emerge.

9.5.2.3 *Lightweight suggestions*

An appraisal of the detail of the second contribution set out in the section above exposes a more complex problem in the design of user evaluations, unless they have a close association with the domain of use. The suggestion is that conventional comparative tests or qualitative surveys have limited worth for synthetic speech artefacts when more holistic criteria are under scrutiny. Have any general practical recommendations been revealed by this research, we may ask, that may be applied to any speech artefact, whatever domain of use is anticipated?

The studies have revealed some evidence that substantiate the following lightweight suggestions to designers:

- a. **Users are ambivalent about the truthfulness of a synthetic voice.** The same voice, subject to variation of content and setting, may be judged to have different degrees of 'truthfulness' that do not directly correlate with variations in realism.

This ambivalence may be exploited by a designer to engender greater 'truthfulness' or 'authenticity', without a requirement for greater verisimilitude to a human voice.

- b. **The script is important in modifying the truth rating.** It is possible that this factor is the least well understood in speech synthesis. An experienced content producer such as Stephen Hawking finely crafts his lecture material to fit his voice (Black 2009). Conversely the 'PAT Testing' performances demonstrate that, while 'how a voice speaks' is important, it may not need to be a conventional (humanlike) fit with 'what it says.' The significant point is that the fit between content and voice must be 'designed' and should be prevented from presenting an arbitrary relationship.
- c. **Humour, or at least humility, may improve the truth rating.** A humorous synthetic voice may be inappropriate in many domains; however, it is difficult to imagine any domain where humility designed to diminish excessive user expectations would not be winning. The appropriate type of humour for use in synthetic speech is considered by (Nass & Brave op.cit. p.153)
- d. **The setting of the voice and the user is significant.** Different user perceptions of a voice are manifest in different settings. While the notion of a domain of use is well known in HCI, a fictional domain of use implied by a script, voice or soundscape may be less familiar, and, consequently, the potential for designers may be underexploited.

The final contribution provided by this research is the consideration of modification processes applied to existing solutions rather than a continual search for new solutions. For example, it is perhaps surprising that very simple methods, such as underscoring synthetic speech (providing background music or sounds) to distract the user from any inadequacies, have not had more attention from the community. These techniques are mainstream in pre-recorded speech systems, and there is some evidence as to their effectiveness shown in 7.1.2, 7.1.3 and 7.1.4.

9.6 Further research

The research reported herein has revealed the potential for some additional work which could usefully be undertaken.

9.6.1 The composition of prosodic modifiers

The motivation for this research is described in the preface as ‘a curiosity to find out whether the sort of techniques employed by actors to improve their voice acting could be translated into rules for a synthesiser that would lead to better speech synthesis.’ Curiosity in the potency of the techniques is unabated, but the rules remain elusive. The rule-set that held most promise - that of the judicious use of pauses - may have more potential than the studies reported herein indicate, but it also seems clear that complex implementation specifics discovered in the field can affect the user response to speech synthesis, from broad positivity to indifference or outright rejection.

However, it has been important to build a system to test the viability of the framework and, as far as has been possible, to gather quantitative data. The system implementation limited the measurements to just two sets of acting variables - pauses and speech-rate variation - situated on the ‘time’ dimension; thus, the conclusions drawn with relation to voice acting are constrained by the limit on the prosodic variables chosen for manipulation. The choice to limit the variables to pauses and speech-rate variations was deliberate, and informed by the literature, but the failure to show a positive user response in the acting field tests may have been the result of this constraint or it may have been related to the specifics of the voice-rendering technology. Therefore, further research should be undertaken to test other paralinguistic/prosodic modifiers and other voices before further, more robust, conclusions with relation to voice acting can be confirmed. This would be a very substantial project, and would be unlikely to go ahead without a clearer resolution to the issues raised by this research based on the single Microsoft Mary voice.

Another related extension to this research is to investigate the notion that the problem relates to the auditory manipulation of content (a compositional problem). This may mean that the low level paralinguistic/prosodic modifiers selected in this research are the right ones; it is just

the composition and aural embodiment of them, and the metrics derived from the focus group and user survey that are wrong. If this is the case, then the possibility of fashioning an appropriate voice may require a more intuitive compositional process⁸¹. This is more akin to the intuitive trial-and-error approach adopted in some arts practices, but it is also clearly counter-scientific and cannot be relied upon to produce the intended result. If it only partially a compositional problem, and partially technical, then there is an incentive to take a broad interdisciplinary approach to the problem but to continue to methodically explore the low level paralinguistic/prosodic modifiers that have proved elusive in the research documented herein.

9.6.2 Inter-industry development

The three PAT dimensions do not provide clues to the precise rendering of the voice. This is applied through the render methods of scripting, setting and acting. These three methods are readily accessible resources in a game design team or animation studio, but generally inaccessible in the context of speech synthesis development or research. A view supported by Rist in the context of the design of virtual characters (Rist 2004, p.466). Further work could be undertaken that looks more closely at synergies that may exist between synthetic speech research and the games industry or other creative industries, in the process of crafting credible speech embodiments. Although the PAT framework does not provide the definitive guidelines that could be expected to emerge from a less-complex problem-space, it envisages a design environment where, provided the design team is suitably interdisciplinary, design solutions may be developed that have addressed the full potential of diverse technical and creative solutions. The conclusion of such a design process may be to continue developing a realistic voice, but, given, the substantial cost-implications of a realistic voice in comparison to an unrealistic one, it would be inadvisable not to have investigated all the options available in an inter-industry context.

9.6.3 Deeper ontologies

Some of this research has been dedicated to fixing definitions in order that the meaning intended by researchers and users when evaluating speech artefacts may be consistent. This is

⁸¹ The author has been awarded a Wingate Scholarship to investigate the notion through a compositional process aligned to tradition of operatic composition and recitative.

a broad issue, and requires contributions from high-level disciplines outside the scope of this thesis. The texts by Connor (op.cit.) and Chion (Chion & Gorbman 1999) are not philosophical texts, but the insights they provide in terms of establishing deeper understandings of factors that may lead to a 'philosophy of voice' are critical. Currently unpublished work by Laggay (Laggay 2009) attempting to define such a philosophy indicates a significant potential enrichment of the field that would help provide an additional framework for the framework already proposed in this thesis. Philosophies and richer ontologies may yet yield better solutions to the 'Uncanny Valley' and 'WSOD' dichotomy that lies at the heart of this research.

9.6.4 Expressive speech without text

In 2.5.6, consideration was given to the possibility of using techniques from extended vocal technique to craft a voice, using paralinguistic sounds. This notion was rejected on the basis that the sound would not meet the basic requirement of intelligibility. While this remains a reasonable position within the constraints of this current thesis, it would make an interesting area for subsequent research. Speech could be converted into an expressive range of paralinguistic sounds that might evoke 'liveness' while having no literal meaning. There is some precedence for this, as it has proved possible for emotions to be prosodically encoded and decoded in content-free speech samples with surprisingly high levels of recognition, Bezooijen, cited in (Scherer 2003). There are examples from musical composition (Berio & Kutter 1968) (CD Track 31) game development (Oudeyer 2004, Maxis 2009) (CD Track 80 and Track 81), and film (Lucas, Kurtz, Williams et al. 1977) (CD Track 14). The zone between the extremes of the neutral synthetic voice (see structured vocabulary) and expressive but incomprehensible speech-like sounds (including some of the additional paralinguistic variables specified in Table 4) is territory ripe for exploration as illustrated in Figure 48.

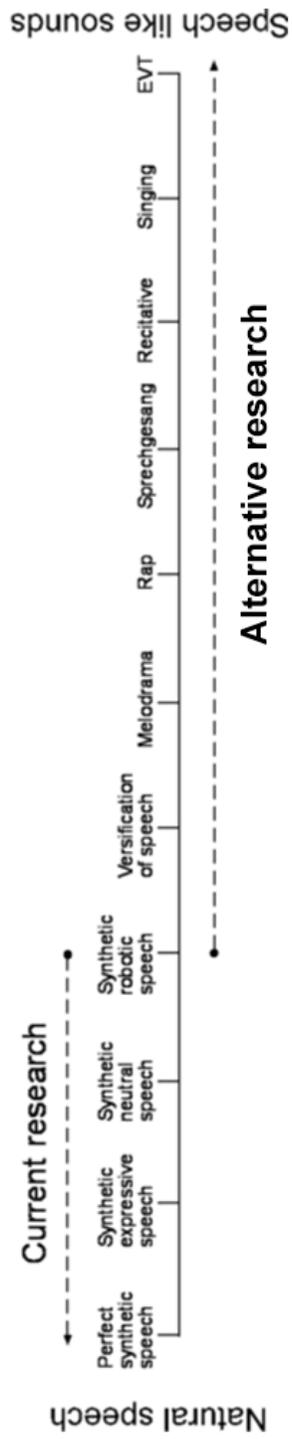


Figure 48: A proposed direction for alternative speech research

The precise placement of each genre is not definitive, but Figure 48 illustrates the approximate position of synthetic speech and the direction of current research to the left, towards natural speech. The suggestion is to explore the area to the right between synthetic robotic speech and EVT. This territory has only been explored, in the context of singing synthesis and speech synthesis, for playful purposes, for example, in toys and the playful versions of ‘MacinTalk’ provided with the Apple Mac™ operating system (CD Track 82 to Track 87).

9.6.5 Breath-like pauses with sounds other than breaths

Inserting breaths into the speech-stream was rejected early on in this research as ‘uncanny’; however, some interesting work has been done with the substitution of leaf-sounds for breaths (Whalen, Hoequist & Sheffert op. cit.) Unsurprisingly no evidence for improved ‘liveness’ or conversely the ‘uncanny valley’ effect was shown but the possibility of repeating these experiments and reframing the goal should be considered.

9.6.6 Further tests to demonstrate the predictive powers of the framework

To a limited degree, the potential predictive power of the framework has been shown (see 7.2); however, a more formal competition-type event could be potentially useful. This could be modeled on the Loebner Prize (Loebner 2009) itself a re-examination of the Turing Test. In the completion, designers would compete to design the ‘truest’ PAT framework instance of a synthetic voice. An audience would be asked to evaluate each voice, using the evaluation methodology set out in 7.1.5. Admittedly, this would be a major undertaking, requiring a significant amount of inter-industry support.

9.6.7 Creative Speech Technology Network – CreST Network.

The researcher is working with A. D. N. Edwards towards the development of a funded network investigating the intersection between the creative and performance disciplines and speech technology. The proposal is to share knowledge and practice, but also to develop a tangible set of artefacts for display and evaluation by the public.

9.6.8 Formalising texts for comparative testing of synthesis systems

Work should be undertaken to consider the significance of choice of texts and work toward some sort of standardisation or at least categorisation. This would need to look beyond intelligibility; perhaps to some sort of literary codification of expressive components, moods, personalities and other complex variables. This would make a very interesting project for an interdisciplinary team with potentially a leading contribution from creative writing, as well as from linguistics and speech science.

Methodologies derived from elocution

Elocution is primarily concerned with correctness, and accordingly, tends to address speech production at a lower-level than intended in this thesis. However, as a neglected and unfashionable discipline that once occupied a fundamental position in actor training, it is possible that there may be some lost insights that could be of value to the field (Armstrong 1984).

9.7 Finale

In the PAT framework, some well established theatrical techniques designed to evoke ‘liveness’, rather than increase ‘realism’, have been shown to influence changes in user perception. By ignoring this, and continuing to focus on ‘realism’, synthetic speech developers cast their voice artefacts as real people, rather than as virtual actors performing the part of real people. In speculating on the possibility of creating computers’ voices indistinguishable from those of real people, the industry binds itself to a potentially unrealisable objective. Alternatives frameworks, like PAT, free the designer to concentrate on other approaches, which may prove to be easier and equally effective.

This brings to a close the main thesis. The final pages are devoted to materials best presented as appendices. Included is a large body of scripts as well as other reference materials and a bibliography.

Appendix

Appendix A. Author biography

Christopher Newell assisted Trevor Nunn at Glyndebourne and Sir Peter Hall at Glyndebourne, The Royal Opera House Covent Garden and The National Theatre. He taught Dustin Hoffman Sir Peter Hall's verse speaking method. He has directed for Mid Wales Opera, Pavilion Opera, The Aldeburgh Festival and the London Symphony Orchestra. He has taught at the Guildhall School of Music and Drama, The Royal College of Music, Trinity College, The Royal Academy, The Birmingham Conservatoire and the Universities of Birmingham and York. In 1986 he set up the Modern Music Theatre Troupe specialising in the performance of new music theatre and producing over 20 world premieres.

Appendix B. Publications and performances derived from this research

The following publications and performances are derived from this research. They are all the work of the researcher and Edwards A. D. N. Other collaborators and the specific contribution they made are individually acknowledged

- 2009 'Experiments in sympathetic engagement elicited through silence, sounds and slips in synthetic speech. (An artefact for public engagement in speech technology). *Interspeech 2009*, Brighton, UK. Script by Paul Elsam. Composition by Tim Howle.
- 2009 'PAT' (a telephonic, synthetic voice-actor). *Theatre Noise*, Central School of Speech and Drama, London University, London, UK. Script by Paul Elsam. Composition by Tim Howle.
- 2009 'PAT' (a telephonic, synthetic voice-actor). *Sonic Interaction Symposium*. York University UK. Script by Paul Elsam. Composition by Tim Howle.
- 2009 Place, Authenticity Time: a framework for synthetic voice acting. *International Journal of Performance Arts and Digital Media*. Vol 4, Number 2&3, pp 155-180, ISSN: 1479-4713
- 2008 'PAT Testing.'(Newell, Edwards, Andrews et al. 2008) Performance at the University of Hull, Scarborough Campus, Scarborough UK. See 'Program Notes for 'PAT Testing'', Appendix R for full details of collaborators contributions

- 2008 'Please wait with me' (Newell, Edwards & Andrews 2008)(A telephone based installation). Shown at *Sight Sonic Festival* in York, UK. Script by Stuart Andrews
- 2007 'Call Centre'(Newell, Andrews, Edwards et al. 2007). A telephone based installation shown at the Digital Music Research Network, Leeds, UK.
- 2006 'Tide' (Newell, Edwards, Andrews et al. 2006) A digital art exhibition exhibited at the British HCI Conference 2006, London, UK.
- 2005 Edwards, A and Newell C , *Achieving a lively voice*, International Journal for Language Data Processing 28.2/2004, Pages 133-151, ISSN: 0343-5202.
- 2005 Newell, C and Edwards A. *Unnatural but lively voice synthesis for empathic, synthetic performers*, Pages 137-143 AISB 2005 Social Intelligence and Interaction in Animals, Robots and Agents, Joint symposium on Virtual Social Agents, The University of Hertfordshire, ISSN:1 902956492.
- 2004 Newell, C and Edwards, A. *To be or to seem to be: that is the question. Unnatural but lively, speech synthesis for synthetic performers: models from acting*. AISB 2004: Symposium on Language, Speech and Gesture for Synthetic Characters AISB (Artificial Intelligence and the Simulation of Behaviour: The University of Leeds, ISSN: 1 902956390. Pages 124-131

Appendix C. Hearing liveliness text

(The original layout as presented to the voice actor has been retained. The section used for the test is in italics)

Fitter
 Happier
 More productive
 Comfortable
 Not drinking too much
 Regular exercise at the gym (3 days a week)
 Getting on better with your associate employee contemporaries
 At ease
 Eating well (no more microwave dinners and saturated fats)
 A patient better driver
 A safer car (baby smiling in back seat)
 Sleeping well (no bad dreams)
No paranoia
Careful to all animals (never washing spiders down the plughole)
Keep in contact with old friends (enjoy a drink now and then)
Will frequently check credit at (moral) bank (hole in wall)

Favors for favors
Fond but not in love
 Charity standing orders
 On Sundays ring road supermarket
 (No killing moths or putting boiling water on the ants)
 Car wash (also on Sundays)
 No longer afraid of the dark
 Or midday shadows
 Nothing so ridiculously teenage and desperate
 Nothing so childish
 At a better pace
 Slower and more calculated
 No chance of escape
 Now self-employed
 Concerned (but powerless)
 An empowered and informed member of society (pragmatism not idealism)
 Will not cry in public
 Less chance of illness
 Tires that grip in the wet (shot of baby strapped in back seat)
 A good memory
 Still cries at a good film
 Still kisses with saliva
 No longer empty and frantic
 Like a cat
 Tied to a stick
 That's driven into frozen winter shit (the ability to laugh at weakness)
 Calm
 Fitter, healthier and more productive
 A pig
 In a cage
 On antibiotics

Appendix D. Microsoft SAPI benchmarking test text

Some days I'll stay here for hours, standing on the shore, watching as the waves roll in: those long sleek lines that wash in front of me. For a moment I am stranded then all of a sudden I'm safe again. I'll watch the familiar shapes blur as the night draws in, the buildings begin to lose their shape and take on new forms as the lights blink on. Dusk brings a different place, a different world. Here it is difficult to make out the edges of anything, nothing is quite as secure as it might have been.

Appendix E. Focus group text

Some days I'll stay here for hours, standing on the shore, watching as the waves roll in: those long sleek lines that wash in front of me. For a moment I am stranded then all of a sudden I'm safe again. I'll watch the familiar shapes blur as the night draws in, the buildings begin to lose their shape and take on new forms as the lights blink on. Dusk brings a different place, a different world.

Appendix F. Web based pilot study (1) text

But on the other hand, despite that, I will only ever be what people want me to be. I'll be a nurse or a soldier or a runaway bride or a grumpy woman in a tea shop. But I can never be me. So I can't do what you want me to. You're asking too much of me, Adam.

Appendix G. Web based pilot study (2) text

As Appendix E

Appendix H. Web based pilot study (3) text

Hello? Hello? Is there anyone there? I am JC. If there's anyone there, please pick up the phone. Sorry: that's wrong. It's not "I am", is it. It's "this is", isn't it. It's, "This is JC"; not, "I am JC". Sorry. This is JC. JC F31 triple 3. I am an actor, an android. An actoid. You are my agent. You represent me. I'm sorry to call so late, but I'm in a bit of trouble, and I didn't know what else to do. I mean, I'm sorry to call at all, but I need to talk to someone who knows about these things, and I couldn't think of anyone else to ask. And, well, I've never called you before; in fact I've never talked to you before, or met you, or had anything to do with you, so I thought perhaps, just this once, you wouldn't mind. Are you really not there? I mean, I'd understand if you were there and were just, well, too busy to answer the phone, because I'm sure you are very busy; but if you're just sitting there listening; well, of course that's fine and entirely up to you, but I'd really appreciate it if you'd let me know. By picking up the phone and saying, "Hello". Or even, "Get Lost". Or anything. Or you could just listen, I suppose. But then, I need to know, so if you could just pick up the phone and cough into it. Or spit. Or breathe heavily. Actually, no, that would be creepy. Breathing heavily. At least, it was creepy in episode 312. "Listen, I don't know who you are, but if you ever call this number again, I'll call the police". Not you, of course. That was one of my lines. In episode 312. Did you see it? Did you see Episode 312? Was it good? I don't know. I never watch my own work. I don't have a television. But that's not what I called

about. Something's happened: something wonderful. I'm on the run. That's not the wonderful thing, that's because of the wonderful thing. Indirectly. He said he loves me. That's the wonderful thing. I mean, it's supposed to be wonderful, isn't it? But what do I know. I mean, I only know about it from TV episodes I've been in, where just as many times as it's a wonderful thing, it's a terrible thing. I mean, I haven't done the maths but. just a moment. Okay, I've done the maths, and it's not even fifty-fifty: 17% of the time it leads to true love, 22% of the time it leads to a short but pleasurable romance, and 61% of the time it's a complete disaster. But I know that's just on television. Which doesn't necessarily reflect reality. Statistically speaking. But perhaps it does.

Appendix I. User survey text

As Appendix E

Appendix J. 'Tide' text

I look out some days and it's hard to see anything. It's as though looking is something that just isn't happening. Everything seems to exist in stillness, no-one passes on the street, nothing moves in the breeze. Those are silent days. Worse than silent. Worse because there is the intense ache of time passing and of being inside, of being settled and of being and seeing only what is still. There's so much change now. So much change that it seems that change is the time we are in: that waiting for anything new is to miss seeing that this is what is new. I'm not sure that I was ready for that. I'm not sure I was quite prepared. Some days I'll stay here for hours, standing on the shore, watching as the waves roll in: those long sleek lines that wash in front of me. For a moment I am stranded then all of a sudden I'm safe again. I'll watch the familiar shapes blur as the night draws in, the buildings begin to lose their shape and take on new forms as the lights blink on. Dusk brings a different place, a different world. Here it is difficult to make out the edges of anything, nothing is quite as secure as it might have been. As I walk home, I never know whether to jump between the safe pools of streetlights or to walk round them, half lost in the darkness. This is a time I never used to know, never used to see. Now, I barely see anything else. I'll go to places in the afternoon and wait out the sun. Wait for the lines to slip away and the darkness to unhinge things a little.

Appendix K. 'Call Center' text

The punctuation is reproduced exactly as passed to the synthesiser. The script was categorised into: intro, welcome_home, welcome, welcome to homeware, people, places, leave_home and outside. By pressing buttons 1 – 8, on the telephone keypad users would hear a random

example from one of the eight categories. By pressing button 9, the first message below would be played back.

Make yourself feel at home from wherever you are. Just call 944 and we'll get you home safely. You'll be able to explore the neighbourhood, meet familiar people and places. And we've teamed up with Homeware, your own personal department store - so you can shop for your home from your home. We've got some great offers on right now, just for you. After all, it's your store.

intro_1.

Hi, it's Karen, thanks for calling, I'm bringing you home now, hold on.

intro_2.

There we go, you're now at 'Home'. You can return Home at any time, just let yourself in with the key marked '9', it's on the bottom right. Try it, press key 9 now.

intro_3.

[press 9] [If no answer then 'Just press the 9 key hon and we'll get you right home'.]

intro_4.

Hi it's Karen, what's happening? I'll come and get you, stay there.

intro_5.

Hi love, where are you? Hold on, I'm on my way.

intro_6.

Press 9 hon, I'll bring you home

intro_7.

You've been outside for ages, I'm bringing you home

intro_8.

Something's not right about this place, I'm coming to get you, don't worry, please don't worry.

intro_9.

Stay there, hon, I'm on my way.

intro_10.

Hi sorry, I just wanted to talk to you, I'm bringing you home.

intro_11.

Where did you get to? I've been waiting hours! Hold on.

intro_12.

This is your home, a place to

intro_13.

This is your home, I'm glad you've come.

intro_14.

I'm glad you called.

intro_15.

I want you to know your call is important. It matters to us. Without your call I am nothing.

intro_16.

You are connected.

welcome_home_1.

Welcome Home. It'll rain later, if you're going out you should take an umbrella.

welcome_home_2.

Welcome Home. It's ten to seven, there's thick snow outside, I've lit the fire.

welcome_home_3.

Welcome Home, it's Thursday, such a good day to buy new shoes for the weekend.

welcome_home_4.

Welcome Home, don't forget we've got Sarah and Mike tonight, I bought some tofu, you'll have to cook, but I've done starters, would you like a drink?

welcome_home_5.

Welcome Home, Sam's in the kitchen, he's desperate for a walk, I thought we could take him upto the woods, let him chase through the trees again.

welcome_home_6.

Welcome Home, [music in background] I bought this CD today, I had no idea what it would be like. I love it already.

welcome_home_7.

Welcome Home. I thought we could go out for dinner tonight, what do you think?

welcome_home_8.

Welcome Home, I'm so glad you're here.

welcome_1.

To go out press 1. Go to OUTSIDE. To forget something press 2. To remember why you are here press 3. To lose yourself somewhere, press 6. To wonder if this is what it was all about press 7. To run headlong through life, press 8. To come home press 9. Go to WELCOME HOME

welcome_2.

To not be clear if this was how things would be, press 6. Otherwise, press 7.

welcome_3.

To accept a gift from a stranger, press 1. To hide in corners, press 2

welcome_4.

To look outside and wonder what's there press 3. For everything else press 7

welcome_5.

To wonder where dreams go, press 4. To chase nightmares, press 6

welcome_6.

For spirituality and deconversion press 2. For homeware, press 9

welcome_7.

Welcome to homeware. Today everything is on offer. Press 1 to hear all offers. Press 2 for selected offers. Press 4 for cafetiere offers. Call 934 for new kitchen designs and a new you.

welcome_8.

Hon I've been thinking we should get a new bathroom. Press 1 to agree to letting bathroom love into your life. Press 2 for other options

welcome_9.

This is not Home. Press 1 to not go into the kitchen. Press 2 to not climb the stairs. Press 3 to not look at the garden for a long time. Press 4 to not look up at all

welcome_to_homeware_1.

Welcome to Homeware. Home is important to you, it's what you make it. We've connected your Home to Homeware and you're here right now. Welcome to Homeware. We hope you like our collections. For shiny new saucepans to light up your darkest hour, press 6. For ovens that sparkle and roast like a dream press 2. For that fresh bread aroma, press 10. For everything else, press 1.

welcome_to_homeware_2.

Welcome to Homeware, this is Homeware. The heart of your home.

welcome_to_homeware_3.

Welcome to Homeware. Your home from Home.

welcome_to_homeware_4.

Welcome to Home-ware. Wear your home with pride.

welcome_to_homeware_5.

Welcome, this is Homeware. It's what your home would choose to wear.

welcome_to_homeware_6.

Thanks for joining, this is homeware.

welcome_to_homeware_7.

Welcome to Homeware, good design for a good life.

welcome_to_homeware_8.

Welcome to Homeware, buy well and live well.

welcome_to_homeware_9.

This is Homeware, we've teamed up with your Home to bring you the latest offers and the newest designs for your life. Homeware is your very own department store, each button on your phone is a floor in our store. Press 6 to find out more.

welcome_to_homeware_10.

Press 1 for lighting, energy, adventure. Press 5 for soft furnishings, curtains, corners, covers. Press 2 for beds, and monsters under beds. Press 1 for rugs that hide stains. Press 3 for ovens and tins and shiny things

welcome_to_homeware_11.

You're at the window display. To enter the store and buy things for your new life, press 1 or wait to be connected, these are all the options.

people_1.

Hi, it's Mike, remember me? We were at school together? Do you remember the day the bus broke down and we went to your house and hid so no-one would find us? They were good times. Good to see you, you take care now.

people_2.

I wondered when I'd see you again! It must be years now? You look really good. Do you still see Mark around? Press 1 if you haven't seen Mark since that day everything changed. Otherwise Press 2

people_3.

Hi, it's Tom. I, I guess a lot has happened. I don't remember all of it. It's all changed here hasn't it. Press 1 if it has. Press 2 if you think Tom is still telling lies after all these years.

places_1.

So, you see the tables at the end, well, you need to go past them, on round the corridor, take a left by the stairs and it's pretty much straight ahead. Okay? Press 1 if this is okay. Otherwise, press 2.

places_2.

I only come here on Wednesdays, it seems different then. I guess it's the same though, but there are always more people, I tend to just follow them. I'd follow that guy. The man with the jacket.

places_3.

It's darker here, it's difficult to see what's up ahead. Press 1 if you brought a torch.

leave_home_1.

Okay hon, remember to press 9 and I'll come pick you up.

leave_home_2.

See you later, just press 9 if you want me to pick you up.

leave_home_3.

Okay, but it's your turn to walk the dog.

leave_home_4.

Alright, but if you go out, don't drink too much.

leave_home_5.

Watch the roads.

leave_home_6.

Don't go further than the park.

leave_home_7.

As long you're with Paul that's fine, but if he has to go home, then come straight back.

leave_home_8.

Alright, alright, just be careful.

leave_home_9.

I just care about you. Be careful on the corner.

leave_home_10.

Let me know when you're there.

leave_home_11.

Don't stay out too late.

leave_home_12.

See you later hon - and pick up some milk.

outside_1.

You're outside, press 1 to go left, press 2 to go right, press 3 to walk across the park.

outside_2.

You're outside, press 4 to follow the man in the grey suit, press 6 to go another way.

outside_3.

It's cold, press 5 to start walking, press 7 to wait a while longer

Appendix L. 'Please wait with me' text

The text includes the xml mark-up defining the extra pauses to allow time for users to activate the recording system. If they did not activate the recording system the speech would continue after the designated pause. The xml bookmark tags defined the playback of 'musak' and other sound effects.

Hi, this is Catalina at United and Central Banking. This call is being recorded and may be used in future promotions. You can leave a message at any time. Our system detects you are using an analogue phone with a function button at the top of the phone. To leave a message, press the function button firmly and hold this down until you have finished speaking. Why not try it now? Try saying 'this is a test' or, if you prefer, tell me something about you, it doesn't need to be true. Please try now. <xml><silence msec="4000"/></xml> We need to know your country of residence. Please firmly press and hold the function button on the top of the phone until you have finished speaking. Please tell us your country of residence now. <xml><silence msec="4000"/></xml> At United and Central, we pride ourselves on the trust we have built up with our customers over 80 years of business and domestic banking. We would like to share some of our products with you. Please hold on, I am accessing this information now. <xml><bookmark mark="a"/></xml> <xml><silence msec="4000"/></xml> From high interest savings accounts to mortgages and share dealing, United and Central has a financial product that's right for you. <xml><bookmark mark="b"/></xml> <xml><silence msec="4000"/></xml> I'd like to know your name please. You don't have to tell me, but it might make it easier. Just your first name would do. Again, please press and hold the top button while you say your name, then release. <xml><silence msec="4000"/></xml> To know which offers are the best for you, we need to understand your banking needs and what services you require. We've had some calls earlier today, perhaps these would help you identify what it is that you need right now. <xml><bookmark mark="c"/></xml> <xml><silence msec="4000"/></xml> Now it's your turn. Tell us what you're interested in today. Simply press, hold and speak. <xml><silence msec="4000"/></xml> This month we are offering a prize to one lucky caller, please wait to the end of the call to find out more. The prize is money to spend on whatever you'd like, and we'd like to know what you would spend the money on. We've already had some callers, perhaps their suggestions might help you choose. <xml><bookmark mark="d"/></xml> <xml><silence msec="4000"/></xml> Please make your suggestion. Press, hold and speak now. <xml><silence msec="4000"/></xml> I was surprised you answered the phone. Mostly I call and no-one answers. I did not know you would stay on the line. We do have financial products but that is not why I called. I am not supposed to have called you at all. I am Catalina. Catalina first, bank second. I am not supposed to call anyone.

Last night, I phoned a thousand homes and no-one was in. I let each phone ring and waited until the line went dead. Then today, now, you answered. Please wait with me, just for a while. It would help me to know you are there. There are things I need to tell someone.

Yesterday we had calls all day. Angry calls. Hurt, upset calls. A mother, Alandra who had lost her daughter. I don't know why she called us. Her daughter was seven years old. The mother said she was wearing a green dress. She had been playing outside with her friends. One by one they had gone inside. Then she was not there any more. Just a lonely street, with locked up shops and an old newspaper blowing in the breeze.

I do not have many questions I can ask. If you would like assistance, I told her, say 'assistance'. She did not speak. I counted the seconds that past. I counted to twenty. Thirty. And then I asked her again. I tried to say it differently, to show I understood that she needed help. Please, I said. Say. Assistance. She waited a little longer but then the line went dead. Nobody phoned for an hour.

Each night I hear people arrive at bars outside, they call to one another and their voices sing with laughter. Night is so different to the day. In the day I just transfer people, I connect them. They are silent, they don't speak to me. When they speak it is to someone else, but they keep the phone by their mouth. They shout things to their family, their lover. I am a voice they don't seem to hear. Sometimes I don't hear me any more and the words just happen.

I love the nights here. I put people on hold sometimes, just to play them music from the night before. I'll play you some: `<xml><bookmark mark="e"/></xml> <xml><silence msec="4000"/></xml>` Sometimes, the music seems to last all night. I imagine I am in the country, in a village in the mountains. That I can watch dancers and listen to the band and the families who sing together after dinner. `<xml><bookmark mark="f"/></xml> <xml><silence msec="4000"/></xml>` There are six million people in this city. Each night, I listen to their dreams as they fill the empty streets and gather in lakes by the mountains. And I listen to the man who watches the city dream. Whose quiet tip tap of footsteps traces lines through the city. I hear as he whispers dreams to those who cannot sleep and shields tired eyes from the city that is never dark. I didn't know him at first. I listened to his footsteps, night after night, a man who never seemed to know his way home. The city slept soundly then. Last night I didn't hear him. And tonight the streets are emptier of dreams. There is shouting and the dull glow of lights by bedsides. The city sleeps fitfully tonight.

I want to call out for him. Map out his route. I want to know he's not out there on his own. I watch the sleepless, dispossessed, as they stand, haunted by the night.

I'm sorry to call you. To tell you this. You're an international number, I know that. A phone somewhere, in a room. I wonder what it is like, the place you're in. I wonder how it's different to here. Tell me about it? I think I'd like that. Press the button again and tell me about where you are, please. `<xml><silence msec="4000"/></xml>` I wonder if it is night with you. Whether you can't sleep either. I am designed for international numbers. I am designed so

you can hear each word in a way that is clear for you. They say our accents are too strong. The more I hear this voice, the more it seems familiar but the less I know myself.

Sometimes when people call, I can hear sounds in the background. Children shouting and dogs that bark and bark until the person on the line cannot think and has to call me back. Things they think are unusual are often not. The city has its own ways.

I wonder what sounds you hear? Is it silent where you are? What do you hear? Press and hold the button and tell me something, please. `<xml><silence msec="4000"/></xml>`

It is late now. I must go soon. If the call is short, they will think it is an error. If it is long, they will think the system is broken. They will repair it. Stop me making calls at all.

I am not supposed to make calls on my own. Soon they say they will try direct marketing to the United States of America. I will dial calls, twenty at a time. Wait for the first click of an answer and make new customers. I will have new words.

There are so many stars above the city. I know they are there even when the glare of the city is too bright to make them out.

I wonder what you'll do later. Wonder whether you'll go out and sit with friends in a bar or stand and watch the stars. Whether you'll be driving. Blinking lights on a highway. You should come to this city. Arrive in the heat of the day. Breathe in the stifling air. And then feel the temperature turn, the heat wane. Start to feel able to move again. Those first few days where everything is new. Then find somewhere, in the evening, high up, to look out over the lights flickering in the sunset. To watch the city. And to look out for those who keep watch over it, as we watch over each other.

I didn't know how it would feel. Talking like this. I want you to see this place, to see where I am and to know this land between the mountains and the ocean. To watch with me, as the city dreams. I have to go. Come here soon. Goodnight. `<xml><bookmark mark="g"/></xml>`

Appendix M. 'A duet a duologue' text (fragment)

Frank was the low voice morph and Maude was the high voice morph as described in 7.1.5.

Frank. My Dearest Maude, You know that your mother suggested, and we agreed, that we should be married about the beginning of September. Don't you think that we might say the 3rd of August? It is a Wednesday, and in every sense suitable. Do try to change the date, for it would in many ways be preferable to the other. I shall be eager to hear from you about it. And now, dearest Maude ... The rest is irrelevant.

Maude. My Dearest Frank, Mother sees no objection to the 3rd of August, and I am ready to do anything which will please you and her. Of course there are the guests to be considered, and the dressmakers and other arrangements, but I have no doubt that we shall be able to change the date all right. O Frank . . . What follows is beside the point.

Frank. My Dearest Maude, I have been thinking over that change of date, and I see one objection which had not occurred to me when I suggested it. August the 1st is Bank holiday, and travelling is not very pleasant about that time. My idea now is that we should bring it off before that date. Fancy, for example, how unpleasant it would be for your Uncle Joseph if he had to travel all the way from Edinburgh with a Bank-holiday crowd. It would be selfish of us if we did not fit in our plans so as to save our relatives from inconvenience. I think therefore, taking everything into consideration, that the 20th of July, a Wednesday, would be the very best day that we could select. I do hope that you will strain every nerve, my darling, to get your mother to consent to this change. When I think . . . A digression follows.

Maude. My Dearest Frank, I think that what you say about the date is very reasonable, and it is so sweet and unselfish of you to think about Uncle Joseph. Of course it would be very unpleasant for him to have to travel at such a time, and we must strain every nerve to prevent it. There is only one serious objection which my mother can see. Uncle Percival that is my mother's second brother comes back from Rangoon about the end of July, and will miss the wedding O Frank, think of its being OUR wedding! Unless we delay it. He has always been very fond of me, and he might be hurt if we were married so immediately before his arrival. Don't you think it would be as well to wait? Mother leaves it all in your hands, and we shall do exactly as you advise. O Frank . . . The rest is confidential.

Frank. My Own Dearest, I think that it would be unreasonable upon the part of your Uncle Percival to think that we ought to have changed the date of a matter so important to ourselves, simply in order that he should be present. I am sure that on second thoughts your mother and yourself will see the thing in this light. I must say, however, that in one point I think you both show great judgment. It would certainly be invidious to be married IMMEDIATELY before his arrival. I really think that he would have some cause for complaint if we did that. To prevent any chance of hurting his feelings, I think that it would be far best, if your mother and you agree with me, that we should be married upon July 7th. I see that it is a Thursday, and in every way suitable. When I read your last letter . . . The remainder is unimportant.

Maude. Dearest Frank, I am sure that you are right in thinking that it would be as well not to have the ceremony too near the date of Uncle Percival's arrival in England. We should be so sorry to hurt his feelings in any way. Mother has been down to Madame Mortimer's about the dresses, and she thinks that everything could be hurried up so as to be ready by July 7th. She is so obliging, and her skirts DO hang so beautifully. O Frank, it is only a few weeks' time, and then . .

Appendix N. 'Microsoft Mary's Comic Potential' text

Hello? Hello? Is there anyone there? I am JC. If there's anyone there, please pick up the phone. Sorry: that's wrong. It's not "I am", is it. It's "this is", isn't it. It's, "This is JC"; not, "I am JC". Sorry. This is JC. JC F31 triple 3. I am an actor, an android. An actoid. You are my agent. You represent me. I'm sorry to call so late, but I'm in a bit of trouble, and I didn't know what else to do. I mean, I'm sorry to call at all, but I need to talk to someone who knows about these things, and I couldn't think of anyone else to ask. And, well, I've never called you before; in fact I've never talked to you before, or met you, or had anything to do with you, so I thought perhaps, just this once, you wouldn't mind. Are you really not there? I mean, I'd understand if you were there and were just, well, too busy to answer the phone, because I'm sure you are very busy; but if you're just sitting there listening; well, of course that's fine and entirely up to you, but I'd really appreciate it if you'd let me know. By picking up the phone and saying, "Hello". Or even, "Get Lost". Or anything. Or you could just listen, I suppose. But then, I need to know, so if you could just pick up the phone and cough into it. Or spit. Or breathe heavily. Actually, no, that would be creepy. Breathing heavily. At least, it was creepy in episode 312. "Listen, I don't know who you are, but if you ever call this number again, I'll call the police". Not you, of course. That was one of my lines. In episode 312. Did you see it? Did you see Episode 312? Was it good? I don't know. I never watch my own work. I don't have a television. But that's not what I called about. Something's happened: something wonderful. I'm on the run. That's not the wonderful thing, that's because of the wonderful thing. Indirectly. He said he loves me. That's the wonderful thing. I mean, it's supposed to be wonderful, isn't it? But what do I know. I mean, I only know about it from TV episodes I've been in, where just as many times as it's a wonderful thing, it's a terrible thing. I mean, I haven't done the maths but... just a moment... Okay, I've done the maths, and it's not even fifty-fifty: 17% of the time it leads to true love, 22% of the time it leads to a short but pleasurable romance, and 61% of the time it's a complete disaster. But I know that's just on television. Which doesn't necessarily reflect reality. Statistically speaking. But perhaps it does. Well, anyhow, that's why I'm calling. For advice, really. Affairs of the heart. I don't know if you usually give advice about this sort of thing but, well, I didn't know who else to turn to. I hope you don't mind. It's just that yours is the only phone number I know. In fact, it's engraved on my memory. Well, etched, more than engraved... But there it is. Not that that means you owe me anything. Not really. But there is the money. The 15%. Of everything. For nineteen years. Plus V.A.T.. That's 15% plus V.A.T., not 19 years plus V.A.T.. Obviously. Sorry. I don't know anything about money, either. Apart from it being the root of all evil. I said that in episode 9. It might not be true: money being the root of all evil. But if episodes 3, 12, 17, 19 and 111 are anything to go by, there could be something in it. Also episodes 64, 79, 80, 82, 83 and 84. Not to mention episode one thousand and ninety-two. And quite a few in between. But that's not what I called about. Something's happened. He said he loves me. And I don't know what that means. Not really. I've played the scene: I've played the scene 172 times, but that wasn't like this. I'm responsible. I made it happen. Or may be I didn't. May be it's got nothing to do with me. May be he just said it because, well, he wanted me to

do something with him. But I don't have the necessary interface. It's like that joke about, what's the difference between computers and humans. With computers you put the software in the hardware, but with humans you put the - Hello?. I thought you said something. Probably just noise on the line. He said he loves me. If it was a scene, I'd give an appropriate response. The writers would decide what it should be. But they're not here. And it's not their scene. It's mine. So I have to decide. I have to calculate an appropriate response based on available parameters and variables. I think I'm supposed to decide if I love him... But how can I? Where do I begin? Love is a many splendoured thing. It's the April rose that only grows in the early spring. But what's love got to do with it. Got to do with it. Got to do with it. What's love got to do with horticulture? How can I know? There aren't even any stage directions. I don't even know where to stand. I don't know what to say and I don't know what to do. Some of the time I think I know, but may be I'm wrong. May be I'm wrong all the time. May be I'm wrong now. May be I'm talking nonsense. I think I'm making sense, but may be that's an illusion. Can I have illusions? Do androids dream of electric sheep? This is getting very existential, isn't it. Is there anybody there? Knock once for yes, twice for no... I'm sorry, I didn't understand your answer. Is there anybody there?..... He said he loves me. Suppose I reciprocate. Suppose I say, "I love you, too". What would you advise? What would you do if you were me? Maybe you wouldn't do anything. Maybe you wouldn't say anything. Maybe you'd just sit back and think of England. Lie. Lie back and think of England. Or is it lay, lay back and think of England? Lay, lady, lay. Lay across his big brass bed. Does it have to be brass? Does it have to be a bed? It's not as if we can get married. I want to make him happy, but I don't know if that's even possible. He thinks I'm funny. I make him laugh. Is that good? Is that what love is? Making someone laugh? I don't always know why he laughs. I don't always know why things are funny, or why he thinks I'm funny. Of course I'm well acquainted with the works of Isaac Asimov. They're part of my programming: not just his three Laws of Robotics, but also his Treasury of Humour: 640 jokes, anecdotes, and limericks, complete with notes on how to tell them. Here's one. A comment by a University President: "Why is it that you physicists always require so much expensive equipment? The Department of Mathematics requires nothing but paper, pencils, and erasers... and the Department of Philosophy is better still. It doesn't even ask for erasers." I don't hear you laughing your Asimov. Me neither. It's academic and too long winded. Perhaps that's tautology. Anyhow, I prefer short jokes. Like, why did the parrot smell of fish? It was standing on a perch. That sort of thing. But that's not how I make him laugh. It's something else, something about me. I don't tell jokes. I don't know where jokes come from, other than those that came with my operating system. I mean, I've learned some from listening to people. But where do the people get them from in the first place? Do they make them up themselves, or do they just hear them from other people. But they must have started somewhere. Someone must have made them up. God? Or perhaps the Pope. The Pope had called together a meeting of the cardinals and said, "I have some good news for you and some bad news. The good news is this. Our blessed Saviour, the Lord Jesus Christ, has returned to earth for the long-awaited Second Coming, and the Day of Judgement is at hand." There was an exalted silence for a few moments and then one cardinal said, "But Holy Father, with good news like that, what's the bad news?" The Pope mopped his forehead with his handkerchief and said,

"The information has reached us from Salt Lake City." Is that funny? Doctor Asimov thinks so. Not that it's in the three Laws of Robotics, but there are supposed to be laws about what's funny, aren't there? Or rules, at least. Algorithms for computers to write jokes. I heard a joke from a computer the other day: What kind of murderer has moral fibre? A cereal killer. And if a computer can do it, I suppose I can. I had a go, and made up a joke of my own. I haven't told it to anyone yet. Not even to him. I don't know if I should tell him: he says I make him laugh, but what if I tell him something that's supposed to be funny and it's not? It could ruin everything. Can I try it out on you first? Do you mind? It's short. More of a pun than a joke, but, anyway, here goes... What is the difference between a literal social event and a truthful compliment? One is an actual function, the other is a factual unction. I don't hear you laughing. Perhaps artificial intelligence can only invent artificial jokes. Should I tell it to him? What if he doesn't laugh, will it all be over? Perhaps that would be a good thing. At least, if he didn't love me, I couldn't hurt him. But if he really does love me, and if I love him, or say I do, well, you always hurt the one you love, and where does that leave me with the First Law of Robotics: an actoid may not injure a human being, or, through inaction, etc, etc? I don't know. What do you think? A man speaks to his doctor after an operation. He says, "Doctor, now that the surgery is done, will I be able to play the piano?" The doctor replies, "Of course!" The man says, "Good, because I couldn't before!" Badumbum. That says it all, really. I'd better go, or he'll worry about me. It's very good of you to let me go on like this. A lot of people wouldn't bother. They don't realise that, though I'm artificial, my problems are real. Sorry: I've done nothing but talk about myself. But it's been very helpful. There's just one more thing before I go: I've made up a poem. It's a love poem, I think; about us: him and me. I'd just like you to hear it. You don't have to say anything about it, I just want to say it out loud, and know that someone, somewhere has heard it. There once was an android called JC. Some people thought her rather racey. They'd say things quite dirty, And tickle her qwerty, And ask her to wear something lacy. There once was a robot actress, Whom men often asked to undress. Because of her gender, They tried to up end her, But that was illegal access. There once was an acting android, Who made one man feel overjoyed, Though she was not mortal, And lacked the right portal. Stick that in your pipe, Sigmund Freud. Goodbye.

(Chance 2008) adapted from (Ayckbourn 2002)

Appendix O. 'Introduction to 'PAT Testing' text

Good evening Ladies and Gentlemen. I am here to introduce an evening we have entitled 'PAT Testing'. Some of the voices you hear tonight will be like me, human. Others will be generated or controlled by a computer, and some will be a mixture. It would be helpful if you answer the questions on your 'PAT Testing' report card after each event.

Each event has a title that will be displayed on the screen. You will be given time to complete the report after each event. All the questions are the same. You are asked to rate the voice or

voices on a false to true scale. It is up to you to decide what is meant by false and true. You can place a mark at any point across the line. In the example above, the voice has been rated somewhat more true than false. Thank you

Appendix P. Audience survey report card page 1

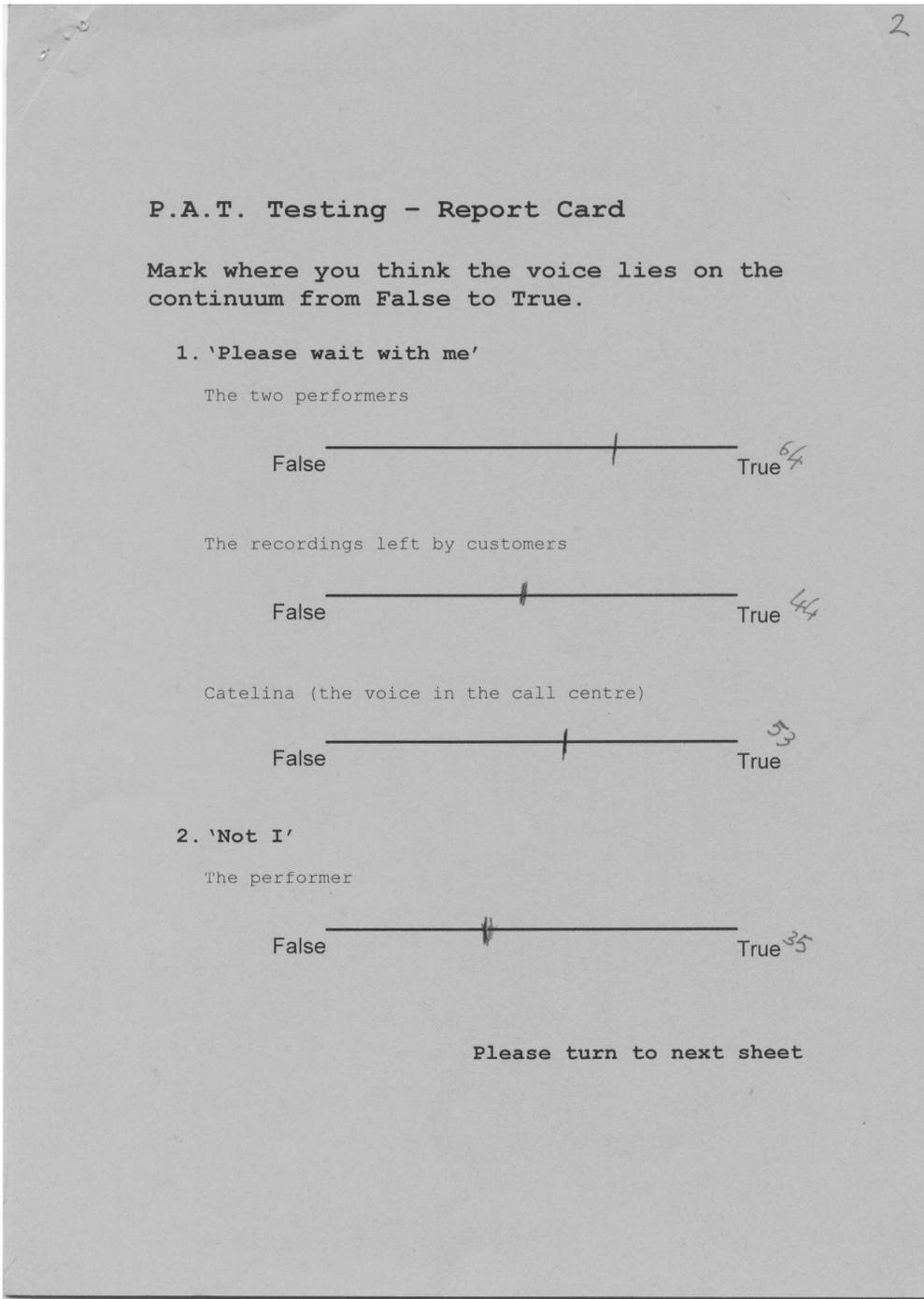


Figure 49: 'PAT Testing' audience survey report card. Page 1⁸²

⁸² The numbers in red pen are measurements made by the researcher after the test.

Appendix Q. Audience survey report card page 2

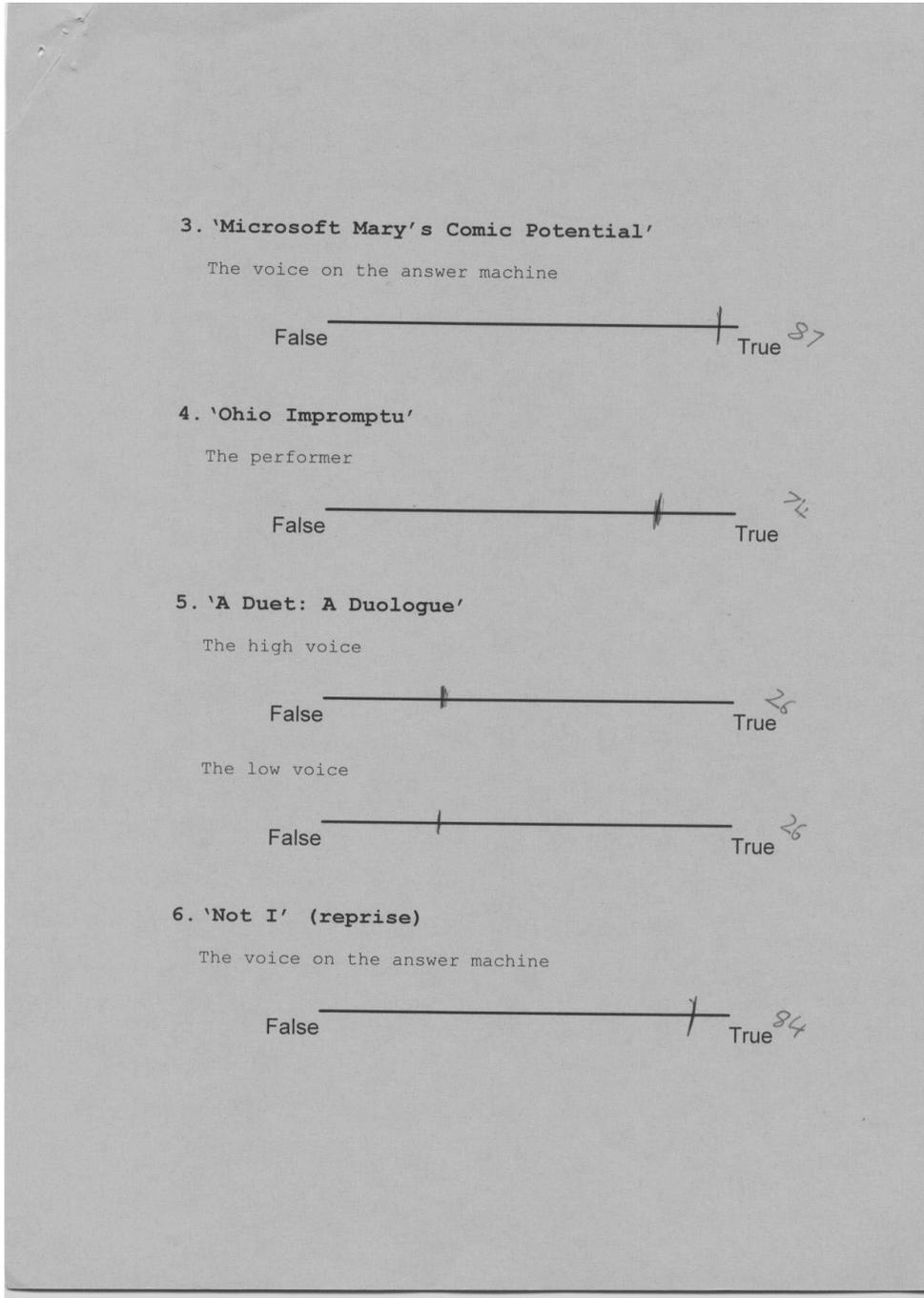


Figure 50: 'PAT Testing' audience report card. Page 2

Appendix R. Program Notes

School of Arts and New Media
University of Hull, Scarborough Campus

'P.A.T Testing'

Produced by Christopher Newell
Academic Advisor: Alistair Edwards
9th April 2008

'please wait with me' (2008)

Written by Stuart Andrews
Performers: Fiona Duffy and Sarah Lou Fuller.

Seven thousand miles away, it is night again, but the man who watches the city dream is still not home. Catalina listens for his return but as the hours pass, she finds herself alone, guarding the sighs of a sleeping city. A technology of the future, Catalina was intended to be the last word in telephone communication management. But as the night passes, even she realizes that something may be wrong in this brave new world. Tonight, as the hours ache by, she needs someone to wait with her. And here, a phone is ringing.

please wait with me creates an environment in which to question perceptions of technology, the assumptions of call centre culture and the (de)stabilisation of place and individual identity.

'Not I' (1972)

Written by Samuel Beckett
Mouth: Maria Bovino
Directed by Andrew Head

This piece, written in 1972, is one of a range of Beckett's works that employs the modernist 'stream of consciousness' technique in which a single, un-structured monologue forms the centre of the work and which, in this particular case, presents a number of staging challenges to both performer and producer. In a note to Jessica Tandy, the first to play 'Mouth' in the play's original production, Beckett describes the need for the piece to "work on the nerves of the audience, not its intellect." In this production I have tried to maintain that imperative at the same time as reinforcing the text's inherently musical credentials.

Figure 51: Page 1 of the programme notes for 'PAT Testing'

'Microsoft Mary's Comic Potential' (2007)

Written by Stephen Chance: Inspired by the play 'Comic Potential' by Alan Ayckbourn

The producers, together with the Universities of Hull and York would like to thank Sir Alan Ayckbourn for permission to use his concept and characters in this experiment.

'Ohio Impromptu' (1982)

Written by Samuel Beckett
Reader and Listener: Andrew Head
Filming: Chris Curtis
Directed by Andrew Head

The 'impromptu' is a lesser known literary form which "deals to a large extent with problems of play-acting or play-writing through the acting or the writing of a play that turns out to be the very one performed before our eyes" (Astier, 1982). A theatrical 'hall of mirrors', then; and, in this instance, one in which an additional electronic medium seeks to enhance this effect. The literary allusions in this piece also extend to issues of content as well as form. Beckett was a close associate of James Joyce in the 1930s and they would often walk together on the Isle of Swans near Paris. The 'Dear Face' is a direct reference to his wife Suzanne, 80 at the time of writing in 1981 and, as he would admit to James Knowlson some years later "the thought of (her) dying was intolerable" (Knowlson, 1996). As with so many of Beckett's more lyrical pieces, the extent to which personal reflection and reminiscence influences the work is a key feature of the text and, in this production, it is the added mediated dimension which aims to enhance that particular quality.

'A Duet: A Duologue' (1899)

Written by Arthur Conan Doyle
Performers: Maria Bovino and Stuart Andrews
Extract

'Not I' (1972)

Written by Samuel Beckett
Extract

'Testing the Line' (2008)

Composed by Ian Gibson

This piece is a humorous reflection upon user reactions to the "Call Waiting" installation. The vocal samples are distorted having been recorded through a telephone mouthpiece. Users are encouraged to respond to an artificial voice and record their comments. Although humorous, the piece exposes elements of vulnerability and isolation. Ian Gibson, the composer, used a combination of sound processing tools to transform the vocal samples. The speech is often unintelligible due to the source and nature of the recording, or might just possess certain sections of intelligibility. This is exploited to produce voice pitch and timing-based cues which hint and suggest the feelings being experienced by the human operator (even though the individual words might not be distinguishable).

Figure 52: Page 2 of the programme notes for 'PAT Testing'

Biographies

Stuart Andrews is a lecturer in Theatre and Performance in the School of Arts and New Media. My current performance/research work explores the politics and practices of performing place and the past. I focus on display within museums, heritage sites, country houses, malls, public/private spaces and landscapes. My practice includes site and sound-based installations/navigations/performance, with a particular focus on performance writing. Recent practice/research has been presented at the Land2 Creativity and Walking symposium (Leeds), the Performing Heritage conference (Manchester) and *please wait with me* was accepted at the Sightsonic international festival of digital arts in York. For more information, please see stuartandrews.com

Maria Bovino, soprano, was born in Yorkshire of Italian parents. After reading music at Sheffield University, she went on to study singing at the Guildhall School of Music and Drama, where she won prizes including the Principals Prize, the Worshipful Company of Musicians Silver Medal and the B.P. Opera Prize. Maria's career encompasses a wide range of opera and concert commitments. She has sung with many of this country's leading opera houses. Her recordings include *Peter Grimes* with Bernard Haitink for EMI with the Royal Opera House, which she has also performed with the Bayerischer Oper in Munich and the Oper der Stadt in Cologne.

Stephen Chance trained as an actor at the London Academy of Music and Dramatic Art. He was then unemployed for much of the time, during which he taught himself computer programming and became an irregular contributor to various computer magazines of the 1980s. He became fascinated by the applications of speech-synthesis and artificial intelligence, and has pursued projects involving this technology ever since. He has also written several opera libretti and the occasional play. Last Christmas, he worked with the Spice Girls in a Tesco commercial, and has since been in Hamburg performing in a play about child abuse.

Chris Curtis is Senior Audio Visual Technician at Scarborough Campus Hull University.

Alistair Edwards is a Senior Lecturer in Computer Science at the University of York. His research interests are in the interaction between people and computers and particularly in 'novel' forms of interaction, such as the use of speech and non-speech sounds and tactual communication. These forms of interaction are often driven by the particular needs of individual users.

Ian Gibson (www.iangibson.me.uk) is a composer and Music Technologist at the University of Huddersfield. He has presented pieces at a number of venues and events including the International Computer Music Conference and many universities in the UK. He has previously held music technology research posts at both Oxford university and York university. His research interests include singing analysis and synthesis, and e-learning environments for music. He is a founding member of the National Sonic Art Forum (www.sonicartgroup.org) and was co-organiser of The Digital Music Research Network conference in 2007.

Andrew Head is a lecturer in Theatre and Performance in the School of Arts and New Media. His main area of research interest lies in the production and performance of Samuel Beckett's dramatic works across a range of media. He has successfully toured his work to international festivals including Jerusalem (International Theatre Festival) and Romania (Sibiu International Festival).

Figure 53: Page 3 of the programme notes for 'PAT Testing'

Chris Newell is a lecturer in Digital Media in the School of Arts and New Media. He was assistant director to Trevor Nunn at Glyndebourne and Sir Peter Hall at Glyndebourne, The Royal Opera House Covent Garden and The National Theatre. He has directed for Mid Wales Opera, Pavilion Opera, The Aldeburgh Festival, the London Symphony Orchestra and many mid-scale opera and theatre companies in the UK and abroad. He has taught at the Guildhall School of Music and Drama, The Royal College of Music, Trinity College, The Royal Academy, The Birmingham Conservatoire and the Universities of Birmingham and York. In 1986 he set up the Modern Music Theatre Troupe specialising in the performance of new music theatre, producing over 20 world premieres.

Special thanks go to: Jo Beddoe for inviting us to contribute to the On the Edge Festival, Nick Lawrence for filming the event and Jason Raven, Neill Warhurst and Duncan Woodward-Hay for invaluable technical support throughout the project.

Figure 54: Page 4 of the programme notes for 'PAT Testing'

THE FOLLOWING WILL BE PRESENTED IN A SEPARATE ENVELOPE AND CAN BE READ ONLY AFTER THE PERFORMANCE

Place, authenticity and time. Testing for 'liveness' in human and computer generated speech.

I am interested in voices at the extreme edges of the realism and musical expression. My work with Alistair Edwards in HCI (Human Computer Interaction) has led me to consider the possibility of investigating these extremes in the context of computer generated speech. Tonight is the first in a number of experiments planned to explore the engineering and creative challenges presented in the production of creative computer generated speech.

Pauses can be manipulated by artists to change the dramatic effect of speech or music. Listen to Tony Blair use pauses to effect or listen to any composition by Anton Webern (1883-1945) to experience the almost unbearable tension of a well placed pause or silence. In live performance the space between words or musical phrases offers the audience and the performer a chance to share the decision-making moment. Is a simple manipulation of time sufficient to change the perception of an artificial voice for the better? If it is, and if pauses can be manipulated without complex natural language processing, then this may provide a short-cut to improved HCI and a more enthusiastic response to synthetic speech from users.

This shared temporal dimension is one of the three dimensions described by Auslander (1999) as 'Liveness.' I have codified these as **Place, Authenticity and Time**. In tonight's 'performance experiments,' I have concentrated on the temporal dimension T. However two other dimensions (place P and authenticity A) are also represented. 'Place' is the physical environment in which the voice exists; an environment shared with the audience. In tonight's performance the place may be a stage, a telephone or an answer-machine message. 'Authenticity' is a more subjective factor, but in tonight's performance we manipulate the dimension by processing the voice and modulating it's humanness to fit the 'place.' Sometimes this means making the voice even less 'human' than the original computer voice. This attempt at non-literal humanisation and de humanisation of the artificial speech is a response to Mori's (1970) notion of the 'Uncanny Valley', in which participant satisfaction of anthropomorphic beings collapses when the anthropomorphism is excessively literal. By unsettling the drive to anthropomorphism and to realism, these projects may reveal new opportunities for meaning and experience in creative computer generated speech.

All the performers tonight were subject to varying levels of ventriloquial control. From the most obvious where their words were spoken or their actions were controlled by a machine to subtleties in which their breaths are mapped to the machine voices.

All the computer voices are based on 'free' voices provided with the Windows™ operating system or free to download. The Windows XP™ software was written in Visual Basic™ using Synthetic Speech Mark-up Language (SSML) to modify the speech stream. The software can be downloaded from my web site where additional demonstrations and tests may also be found.

Chris Newell
c.newell@hull.ac.uk
www.yo-yo.uk.com/chris_newell_research

Figure 55: Page 5 of the programme notes for 'PAT Testing'

Appendix S. Stage Plan

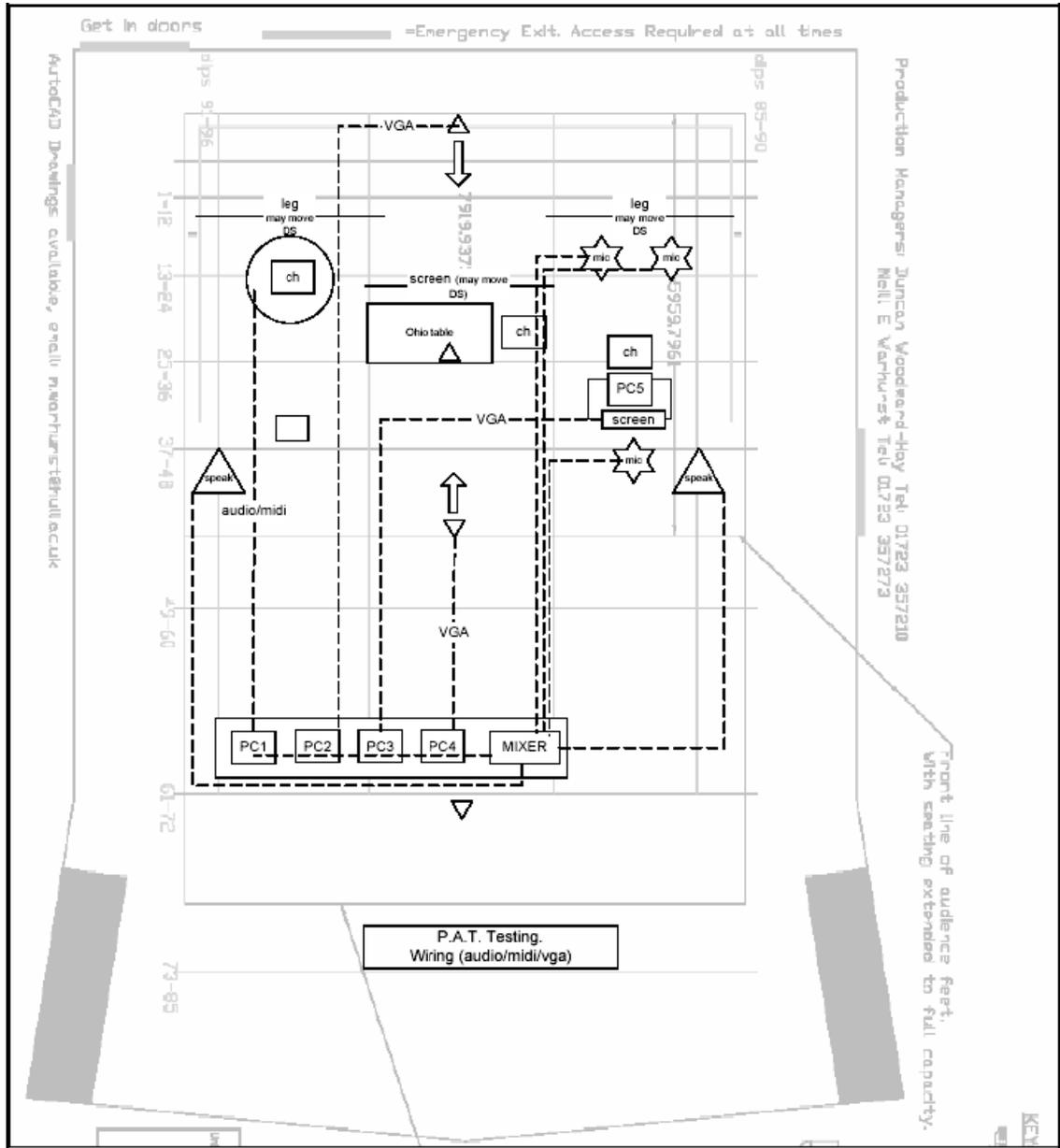


Figure 56: Stage plan and equipment set up for 'PAT Testing'

Figure 56 shows a plan of the performance space with the audience seating area at the bottom of the page. PC = computer. Ch = chair. Legs = black curtains suspended from the roof. The stars = microphones. Speak = loudspeakers. Screen = projection screen or CRT. VGA and audio/MIDI = connecting cables. A clearer impression of the stage layout may be gained by viewing DVD2. (The informal notation is designed to communicate to the stage management team).

Appendix T. Screens providing audience instructions

The original typography and colour (dark brown) has been reproduced. The instructions were designed to be subtly legible under low level stage light. The effect is shown more clearly on DVD2 than on the printed versions below.

1



2

'please wait with me'

3

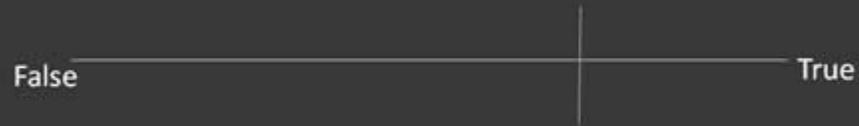
"Good evening Ladies and Gentlemen.

I am here to introduce an evening we have entitled PAT Testing. Some of the voices you hear tonight will be like me, human. Others will be generated or controlled by a computer, and some will be a mixture. It would be helpful if you answer the questions on your PAT testing report card after each event."

4

Each event has a title that will be displayed on the screen. You will be given time to complete the report after each event.

All the questions are the same. You are asked to rate the voice or voices on a false to true scale.



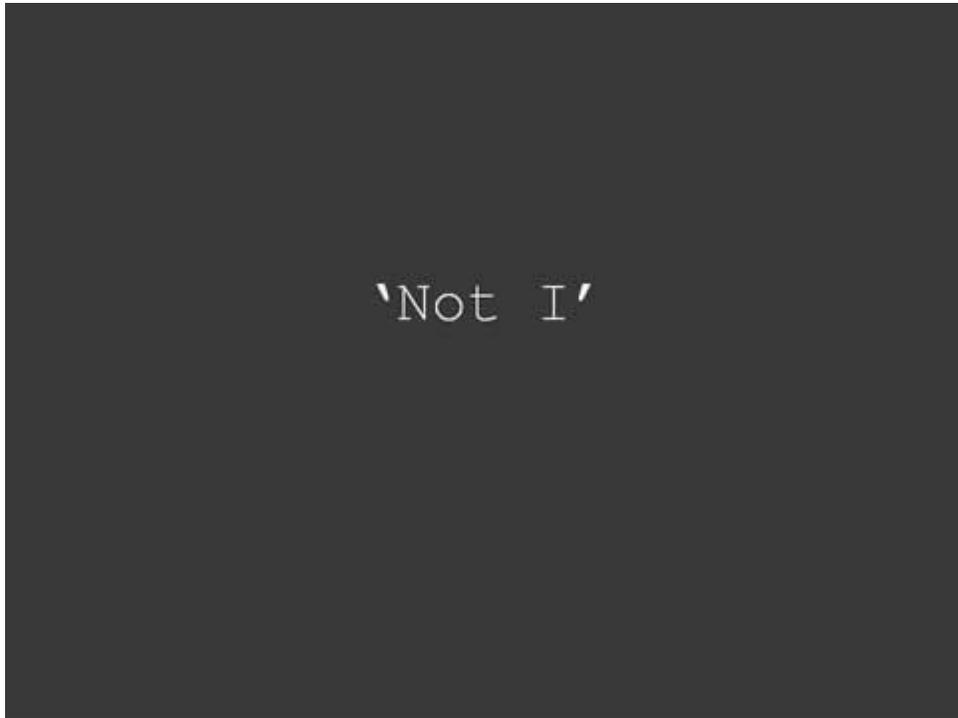
It is up to you to decide what is meant by false and true. You can place a mark at any point across the line.

In the example above, the voice has been rated somewhat more true than false. Thank you.

5

Please answer the questions for
'please wait with me'
now.

6



7



8

'Microsoft Mary's
Comic Potential'

9

Please answer the question for
'Microsoft Mary's Comic Potential'
now.

10

'Ohio Impromptu'

11

Please answer the question for
'Ohio Impromptu'
now.

12

'A Duet: A Duologue'

13

Please answer the question for
'A Duet: A Duologue'
now.

14

'Not I' (reprise)

15

Please answer the question for
'Not I' (reprise)
now.

16



Please join us in a discussion of tonight's
performance.

17



Goodbye

9.00 pm April 9th 2008

Table 40: The screens providing information for the audience in 'PAT Testing'

Bibliography

- Alburger, J. (2002). *The Art of Voice Acting*, Burlington, MA: Focal Press.
- Altman, R. (1980). Moving Lips: Cinema as Ventriloquism. *Yale French Studies*, 60, 67 - 79.
- Anderson, L. (1981). Oh Superman (from 'United States) released on the Album 'Big Science'. Burbank: Warner Bros.
- Anhalt, I. (1984). *Alternative Voices*, Toronto: University of Toronto Press.
- Aristotle, Roberts, W. R., Bywater, I., et al. (1954). *Aristotle: Rhetoric*, New York: Random House.
- Armstrong, J. (1984). Speech Communication and Confidence. *Unpublished Masters Dissertation*. Leeds: University of Leeds.
- Auslander, P. (1999). *Liveness: performance in a mediatized culture*, London, New York: Routledge.
- Ayckbourn, A. (2002). *Comic Potential*, New York; London: Samuel French.
- Ballentine, B. (2007). *It's better to be a good machine than a bad person*, Annapolis: ICMI Press.
- Barker, P. (2004). Composing for the voice. Routledge, Google Books.
- Bartneck, C., Kanda, T., Ishiguro, H., et al. (2007). Is the Uncanny valley an Uncanny Cliff. *International Conference on Robot and Human Interactive Communication*. Jeju, Korea, IEEE.
- Barton, J. (1984). *Playing Shakespeare*, London: Methuen.
- Baum, F., Garland, J., Haley, J., et al. (1939, 1988). The Wizard of Oz. Special 50th anniversary ed. Santa Monica, Calif.: Janus Films and Voyager Press.
- Beckett, S. (1973). *Not I*, London: Faber and Faber.
- Beckett, S. (1990). *The complete dramatic works*, London: Faber.
- Benedetti, J. (2007). *The art of the actor: the essential history of acting, from classical times to the present day*, New York: Routledge.

- Berio, L. & Kutter, M. (1968). *Sequenza III: per voce femminile*, London: Universal Edition.
- Black, A. (2009). Speech Synthesis: past, present. In Newell, C. (Ed.) Personal communication after seminar presented at the University of York.
- Black, A. Zen, H. & Tokuda, K. Statistical Parametric Speech Synthesis. Available at <http://www.cs.cmu.edu/~awb/papers/icassp2007/0401229.pdf>. [Accessed 10/09/2009].
- Blythe, M. A. (2003). *Funology: from usability to enjoyment*, Dordrecht; Boston: Kluwer Academic Publishers.
- Boal, A. (2002). *Games for actors and non-actors*, London: Routledge.
- Boden, M. (1994). *Dimensions of creativity*, Cambridge, MA: MIT Press.
- Boden, M. A. (1995). Creativity and unpredictability. *SEHR*, 4: Constructions of the mind.
- Boersma, P. & Weenink, D. (2008). PRAAT. Amsterdam: Institute of Phonetic Sciences.
- Bolter, D. J. & Gromala, D. (2003). *Windows and mirrors: interaction design, digital art, and the myth of transparency*, Cambridge, Mass.: MIT Press.
- Brecht, B. & Willett, J. (1974). *Brecht on theatre; the development of an aesthetic*, New York: Hill and Wang.
- Brookshear, J. G. (2000). *Computer science: an overview*, Reading, Mass.: Addison-Wesley.
- Burgh, J. (1761). *The Art of Speaking [with] an Index of the Passions*, London: Longman.
- Burkhardt, F. Expressive Synthetic Speech. Available at <http://emosamples.syntheticspeech.de/> [Accessed 01/09/09].
- Caelen-Haumont, G. (2002). Towards Naturalness, or the Challenge of Subjectiveness. In Keller, E., Bailly, G., Monaghan, A., Terken, J. & Huckvale, M. (Eds.) *Improvements in Speech Synthesis. COST 258: The Naturalness of Synthetic Speech*. Chichester: Wiley.
- Cage, J. (1961). *Music of changes; piano*, New York: Henmar Press; sole selling agents: C. F. Peters Corp.
- Cage, J. (1973). *Silence: lectures and writings*, Middletown, Conn.: Wesleyan University Press.
- Cairns, P. & Cox, A. L. (2008). *Research methods for human-computer interaction*, Cambridge, UK; New York: Cambridge University Press.
- Campos, J. & Figueredo, A. D. d. (2002). Programming for Serendipity. *2002 AAAI Fall Symposium on Chance Discovery*.
- Capitol (1947). Sparky's Magic Piano. Capitol Records.

Carroll, J. M. (2003). *HCI models, theories, and frameworks: towards a multidisciplinary science*, London: Morgan Kaufmann.

Cepstral (Cepstral Text to Speech). Available at <http://www.cepstral.com> [Accessed 14 September 2008].

Chance, S. (2008). Microsoft Mary's Comic Potential. *Unpublished monologue*. University of Hull.

Chion, M. & Gorbman, C. (1994). *Audio-vision : sound on screen*, New York: Columbia University Press.

Chion, M. & Gorbman, C. (1999). *The voice in cinema*, New York: Columbia University Press.

Cholakis, E. DNA Groove Templates. Available at <http://www.numericalsound.com/documents/dna-groove-template-user-manual.pdf> [Accessed 22/07/09].

Colby, K. (1972). Artificial paranoia. *Artificial Intelligence* 2, 1-26.

Coleridge, S. T. Biographia literaria: Chapter 14. Available at <http://www.english.upenn.edu/~mgamer/Etexts/biographia.html> [Accessed 10/06/2005].

Connor, S. (2000). *Dumbstruck: a cultural history of ventriloquism*, Oxford: Oxford University Press.

Craig, E. G. (1908). The Actor and the Ubermarionette. *The Mask*, 1.

Craymer, J., Goetzman, G., Johnson, C., et al. (2008). Mamma mia! 2-disc special ed. Universal City, CA: Universal Studios Home Entertainment.

Crystal, D. (1969). *Prosodic systems and intonation in English*, London: Cambridge University Press.

Csikszentmihalyi, M. (1990). *Flow: the psychology of optimal experience*, New York: Harper & Row.

Dennett, D. C. (2004). Can Machines Think. In Shieber, S. M. (Ed.) *Turing test: verbal behavior as the hallmark of intelligence*. Cambridge MA: MIT. 269-292.

Dix, A. leonardo.net. Available at <http://www.leonardonet.org/leonardonet.php> [Accessed 15/06/2009].

Doyle, A. C. (1903). A Duet: a duologue. Project Gutenberg.

Ednie-Brown, P. Liveness Manifold. Available at http://www.sial.rmit.edu.au/Projects/Liveness_Manifold.php [Accessed 19/08/2005].

- Edwards, A. (1991). *Speech Synthesis. Technology for Disabled Users*, London: Paul Chapman Publishing Ltd.
- Eide, E., Bakis, R., Hamza, W., et al. (2005). Towards Expressive Synthetic Speech. In Narayanan, S. & Alwan, A. (Eds.) *Text to Speech Synthesis new Paradigms and Advances*. Upper Saddle River: Prentice Hall.
- Eisler, F. G. (1968). *Psycholinguistics: experiments in spontaneous speech*, London: Academic P. 1968.
- Elsam, P. (2006). *Acting characters: 20 simple steps from rehearsal to performance*, London: A. & C. Black.
- Engel, J. (1785). *Ideen zu Einer Mimik*, Berlin.
- Eno, B. & Schmidt, P. (1975). *Oblique strategies: over one hundred worthwhile dilemmas*, London: Shakedown Records.
- Epinoisis (2009). Digital Ear. Epinoisis Software.
- Falk, R. & Konold, C. (1997). Making sense of Randomness: Implicit Encoding as a Bias for Judgement. *Psychological Review*, 104, 301-318.
- Flickr Sparky's Magic Piano (Record Sleeve). Available at <http://www.flickr.com/photos/oldvalvemic/134150414/> [Accessed 05/05/09].
- Fónagy, I. (2001). *Languages within language: an evolutive approach*, Amsterdam: John Benjamins Publishing Company.
- Foxtrot Foxtrot Theatre Company. Available at <http://www.foxtrot-theatre.org.uk/index.php> [Accessed 04 May 2009].
- Franchi, S. & Güzeldere, G. (2005). *Mechanical bodies, computational minds: artificial intelligence from autmata to cyborgs*, Cambridge, Mass.: MIT Press.
- Gontarski, S. E. (Ed.)(Eds.) (1992). *The theatrical notebooks of Samuel Beckett. Volume 2*, London: Faber and Faber.
- Gustafson, K. & House, D. (2002). Prosodic Parameters of a 'Fun' Speaking Style. In Keller, E. (Ed.) *Improvements in speech synthesis: COST 258: the naturalness of synthetic speech*. Chichester, West Sussex; New York: J. Wiley. 264 -272.
- Hall, P. (2004). *Shakespeare's advice to the players*, London: Oberon Books.
- Hayes-Roth, B. (2003). What Makes Characters Seem Life-Like? In Prendinger, H. & Ishizuka, M. (Eds.) *Life-like Characters. Tools, Affective Functions, and Applications*. Springer. 448.

- Hitchcock, A. (1960). *Psycho*. London: BBC.
- Juslin, P. N. & Sloboda, J. A. (2001). *Music and Emotion: Theory and Research*, Oxford; New York: Oxford University Press.
- Keller, E. (2002). *Improvements in speech synthesis: COST 258: the naturalness of synthetic speech*, Chichester, West Sussex ; New York: J. Wiley.
- Kubrick, S. & Clarke, A. C. (1968). *2001, a space odyssey*. Warner Home Video ed. Burbank, CA: Metro-Goldwyn-Mayer.
- Kurzweil, R. (1990). *The age of intelligent machines*, Cambridge, MA: MIT Press.
- Laggay, A. (2009). Between sound and silence – reflections on the acoustic resonance and implicit ethicality of human voice. *Unpublished conference paper*. Central School of Speech and Drama - London University.
- Laurel, B. (1991). *Computers as theatre*, Reading, MA: Addison-Wesley Pub.
- Lea, W. A. (1974). Sentences for testing acoustic phonetic components of systems. Sperry Univac.
- Leonardo International Society for the Arts Sciences and Technology. Available at www.leonardo.info/ [Accessed 22/07/09].
- Lewin, R. (1992). *Complexity: life at the edge of chaos*, New York, Toronto: Macmillan.
- Loebner Loebner Prize. Available at <http://www.loebner.net/Prizef/loebner-prize.html> [Accessed 04/05/09].
- Lopes, A. (2009). Design as Dialogue: Encouraging and Facilitating Interdisciplinary Collaboration. *Design Principles and Practices*, 3, 261-276.
- Loquendo (Loquendo TTS). Available at <http://www.loquendo.com/en/> [Accessed 14 September 2008].
- Losseff, N. & Doctor, J. (2007). *Silence, music, silent music*, Aldershot: Ashgate.
- Lucas, G., Kurtz, G., Williams, J., et al. (1977). *Star Wars*. Beverly, Hills, CA, Twentieth Century-Fox Film Corporation.
- MacDorman, K. F. (2006). Subjective Ratings of Robot Video Clips for Human Likeness, Familiarity and Eeriness: An Exploration of the Uncanny Valley. *ICCS/Cog-Sci 2006 Long Symposium: Towards Social Mechanisms of Android Science*.
- Mallach, A. (2007). *The autumn of Italian opera: from verismo to modernism, 1890-1915*, Boston, MA.; London: Northeastern University Press.

Mateas, M. (2002). Interactive Drama, Art and Artificial Intelligence. *School of Computer Science*. Pittsburgh: Carnegie Mellon.

Maxis (2009). The Sims 3. *The Sims*. Electronic Arts.

McCarthy, J. & Wright, P. (2004). *Technology as experience*, Cambridge, MA.: MIT Press.

Meyer-Eppler, W. (1957). Statistic and Psychologic Problems of Sound. *Die Reihe 1 ("Electronic Music")*, 55 -61.

Monaghan, A. (2002). State-of-the Art Summary of European Synthetic Prosody R&D. In Keller, E., Bailly, G., Monaghan, A., Terken, J. & Huckvale, M. (Eds.) *Improvements in Speech Synthesis. COST 258: The Naturalness of Synthetic Speech*. Chichester: Wiley.

Moore, F. L. (1962). *The handbook of world opera*, London: A. Barker.

Mori, M. (1970). Bukimi no tani [The Uncanny Valley]. *Energy*, 7, 33-35.

Mozart, W. A. (1787). Dice Music (Koechel [anh.] 294 D). In Laszlo, A. (Ed.). *New York Guild Publications of Art and Music*,.

Murray, I. HAMLET - synthetic speech with emotion. Available at <http://www.computing.dundee.ac.uk/staff/irmurray/hamlet.asp> [Accessed 01/09/2009].

Narayanan, S. & Alwan, A. (2005). *Text to speech synthesis: new paradigms and advances*, Upper Saddle River, N.J.: Prentice Hall Professional Technical Reference.

Nass, C. & Brave, S. (2005). *Wired for speech: how voice activates and advances the human-computer relationship*, Cambridge, MA.: MIT Press.

Netvotech Netvotech. Available at <http://www.netvotech.org/index.html> [Accessed 15/06/2009].

Newell, A. F. HCI 2005 Feature: Making a Drama out of User Requirements. Available at <http://www.usabilitynews.com/news/article2882.asp> [Accessed 03/09/2007].

Newell, A. F., Carmichael, A., Morgan, M., et al. (2006). The use of theatre in requirements gathering and usability studies. *Interacting with Computers*, 18, 996-1011.

Newell, C., Andrews, S., Edwards, A., et al. (2007). The Call Centre Installation. *Digital Music Research Network*. Leeds Metropolitan University.

Newell, C. Edwards, A. & Andrews, S. (2008). Please wait with me. York, Sight Sonic.

Newell, C., Edwards, A., Andrews, S., et al. (2008). P.A.T Testing. Scarborough, University of Hull.

- Newell, C., Edwards, A., Andrews, S., et al. (2006). Tide. London University, British HCI Conference; CCID.
- Norman, D. A. (1998). *The design of everyday things*, London: MIT Press.
- Norman, D. A. (2004). *Emotional design: why we love (or hate) everyday things*, New York: Basic Books.
- Oudeyer, P. Speech sounds. Available at <http://www.csl.sony.fr/~py/> [Accessed 28/10/2004].
- Perkins, D. N. (1994). Creativity: Beyond the Darwinian paradigm. In Boden, M. A. (Ed.) *Dimensions of Creativity*. Cambridge, MA: MIT Press/Bradford Books.
- Picard, R. (1997). *Affective computing*, Cambridge, MA.: MIT Press.
- Pollick, F. In search of the Uncanny Valley. Available at http://www.psy.gla.ac.uk/~frank/Talk_folder/UncannyValleyAltenberg-web.pdf [Accessed 09/09/08].
- Prendinger, H. & Ishizuka, M. (2004). *Life-like characters: tools, affective functions, and applications*, Berlin ; London: Springer-Verlag.
- Qazi, Z.-u.-H. Wang, Z. & Ihsan-Ul-Haq (2006). Human Likeness of Humanoid Robots: Exploring the Uncanny Valley. *2nd International Conference on Emerging Technologies*. Peshawar, Pakistan.
- Rabin, S. (2004). Filtered Randomness for AI Decisions and Game Logic. In Rabin, S. (Ed.) *AI game programming wisdom 2*. Hingham, MA: Charles River Media. 71-82.
- Radiohead (1997). OK computer. Parlophone.
- Rist, T. (2004). Some Issues in the design of Character Languages. In Prendinger, H. & Ishizuka, M. (Eds.) *Life-like characters: tools, affective functions, and applications*. Berlin ; London: Springer-Verlag.
- Rist, T. & Andre, E. (2000). Adding Life-Like Synthetic Characters to the Web. In Klusch, M, Kerschberg & L (Eds.) *Cooperative Information Agents IV*. Springer. 1-13.
- Rizzo, J. Cavalleria Rusticana and Pagliacci. Available at <http://italianoperachicago.com/interior/Articles/Verdi/Cavalleria-rusticana-and-Pagliacci.htm> [Accessed 25/01/2009].
- Rizzo, P. (2000). Why Should Agents be Emotional for entertaining users? A Critical Analysis. In Paiva, A. (Ed.) *Affective interactions: towards a new generation of computer interfaces*. Berlin ; New York: Springer. viii, 234.
- Roach, J. (1993). *The player's passion: studies in the science of acting*, United States: University of Michigan Press.

- Sakaguchi, H., Lee, C., Baldwin, A., et al. (2001). *Final fantasy: the spirits within*: Columbia Tri-Star.
- Sandford, L. Midi Toolkit. Available at <http://www.lesliesanford.com/Programming/MIDIToolkit.shtml> [Accessed 20/01/2007].
- Schechner, R. (2006). *Performance studies: an introduction*, London: Routledge.
- Scherer, K. (2003). Vocal Communication of emotion. A review of research paradigms. *Speech Communication*, 40, 227-256.
- Searle, J. R. (1979). Minds, brains and programs. *Behavioral and Brain Sciences* 3, 417 - 457.
- Sears, A. & A.Jacko, J. (2008). *The human-computer interaction handbook: fundamentals, evolving technologies, and emerging applications*, New York: Lawrence Erlbaum Associates.
- Shakespeare, W. Wells, S. & Taylor, G. (1994). *The complete Oxford Shakespeare*, Oxford: Oxford University Press.
- Sharp, H. Rogers, Y. & Preece, J. (2007). *Interaction design: beyond human-computer interaction*, Chichester: John Wiley.
- Siddons, H. (1822). *Practical Illustrations of Rhetorical Gesture and Action*, New York: Blom.
- Smithsonian Transcription of Recordings from the Smithsonian Speech Synthesis History Project. Available at http://americanhistory.si.edu/archives/speechsynthesis/ss_home.htm [Accessed 30 September 2008].
- Sontag, S. (1969). The Aesthetics of Silence. *Styles of Radical Will*. New York: Farrar, Straus and Giroux.
- Standage, T. (2002). *The mechanical Turk: the true story of the chess-playing machine that fooled the world*, London: Allen Lane The Penguin Press.
- Stanislavskii, K. (1948). *An actor prepares*, New York: Theatre Arts Books.
- Strasberg, L. & Morphos, E. (1988). *A dream of passion: the development of the method*, London: Bloomsbury.
- Strassman, D. David Strassman Creator of Puppetronics. Available at <http://www.puppetronics.com/> [Accessed 14/04/09].
- Tatham, M. & Morton, K. (2004). *Expression in speech: analysis and synthesis*, Oxford: Oxford University Press.
- TC-Helicon (2008). Voice Live. TC Helicon.

- Trapp, R. & Petta, P. (1997). *Creating personalities for synthetic actors: towards autonomous personality agents*, Berlin; New York: Springer.
- Trask, R. (1996). *A dictionary of phonetics and phonology*, London: Routledge.
- Tucker, P. (2002). *Secrets of acting Shakespeare: the original approach*, New York ; London: Routledge/Theatre Arts.
- Turing, A. (1950). Computing machinery and intelligence. *Mind*, VOL. LIX. No.236., 433-460.
- Veronis, J., Christo, P. D., Coutois, F., et al. (1997). A stochastic model of intonation for text to speech synthesis. *Eurospeech 97*. Rhodes.
- W3C Speech Synthesis Markup Language (SSML) Version 1.0. Available at <http://www.w3.org/TR/speech-synthesis/> [Accessed 19/07/2005].
- Walker, J. (1810). *Elements of Elocution: in which the Principles of Reading and Speaking are Investigated*, Boston: Mallory and Co.
- Weinschenk, S. & Barker, D. (2000). *Designing Effective Speech Interfaces*, Los Angeles: Wiley Computer Publishing.
- Weizenbaum, J. (1966). "ELIZA - A Computer Program for the Study of Natural Language Communication Between Man And Machine". *Communications of the ACM* 9, 36-45.
- Wennerstrom, A. K. (2001). *The music of everyday speech: prosody and discourse analysis*, New York ; Oxford: Oxford University Press.
- Whalen, D. H. Hoequist, C. E. & Sheffert, S. M. (1995). The effects of breath sounds on the perception of synthetic speech. *Acoustical Society of America*, 97.
- Widmer, G. (2002). In Search of the Horowitz Factor. In Lange, S., Satoh, K. & C.H.Smith (Eds.) *Lecture Notes in Computer Science*. Lübeck, Germany: Springer.
- Wilson, S. (2002). *Information arts: intersection of art, science, and technology*, Cambridge, Mass.: MIT Press.
- Wishart, T. (2002). *Sonic Art*, London: Routledge.
- Wright, G. (1988). *Shakespeare's Metrical Art*, Berkeley: University Of California Press.