

Understanding the Phonetics of Neutralisation:

A Variability-Field Account of Vowel/Zero Alternations
in a Hijazi Dialect of Arabic

Mariam Musellim Almihmadi

Department of Linguistics
University College London

Thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy

UCL

2011

DECLARATION

I confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Mariam M. Almihmadi

ABSTRACT

This thesis throws new light on issues debated in the experimental literature on neutralisation. They concern the extent of phonetic merger (the completeness question) and the empirical validity of the phonetic effect (the genuineness question). Regarding the completeness question, I present acoustic and perceptual analyses of vowel/zero alternations in Bedouin Hijazi Arabic (BHA) that appear to result in neutralisation. The phonology of these alternations exemplifies two neutralisation scenarios bearing on the completeness question. Until now, these scenarios have been investigated separately within small-scale studies. Here I look more closely at both, testing hypotheses involving the acoustics-perception relation and the phonetics-phonology relation.

I then discuss the genuineness question from an experimental and statistical perspective. Experimentally, I devise a paradigm that manipulates important variables claimed to influence the phonetics of neutralisation. Statistically, I re-analyse neutralisation data reported in the literature from Turkish and Polish. I apply different pre-analysis procedures which, I argue, can partly explain the mixed results in the literature.

My inquiry into these issues leads me to challenge some of the discipline's accepted standards for characterising the phonetics of neutralisation. My assessment draws on insights from different research fields including statistics, cognition, neurology, and psychophysics. I suggest alternative measures that are both cognitively and phonetically more plausible. I implement these within a new model of lexical representation and phonetic processing, the Variability Field Model (VFM). According to VFM, phonetic data are examined as jnd-based intervals rather than as single data points. This allows for a deeper understanding of phonetic variability. The model combines prototypical and episodic schemes and integrates linguistic, paralinguistic, and extra-linguistic effects. The thesis also offers a VFM-based analysis of a set of neutralisation data from BHA.

In striving for a better understanding of the phonetics of neutralisation, the thesis raises important issues pertaining to the way we approach phonetic questions, generate and analyse data, and interpret and evaluate findings.

To John Harris (my supervisor), Yi Xu (my second supervisor), Moira Yip, Neil Smith, Francis Nolan (my external examiner), and Andrew Nevins (my internal examiner)

Table of Contents

DECLARATION	2
ABSTRACT	3
List of Tables	8
List of Figures	10
1 Introduction.....	14
1.1 Prelude.....	14
1.2 Towards a Better Understanding of the Phonetics of Neutralisation	14
1.3 Neutralisation in Laboratory Studies	19
1.3.1 Background.....	19
1.3.2 Neutralisation and the Phonetics-Phonology Relation.....	21
1.4 Structure of the Thesis.....	26
2 Variability in the Phonetics of Neutralisation.....	28
2.1 Introduction	28
2.2 Manifestations of Variability in the Phonetics of Neutralisation.....	32
2.3 Approaches to Variability in the Phonetics of Neutralisation.....	38
2.3.1 Overview	38
2.3.2 Either Complete or Incomplete.....	42
2.3.3 Only Complete	46
2.3.4 Only Incomplete	49
2.3.5 Flexible Models	57
2.3.6 Summary.....	62
2.4 Conclusion.....	63
3 The Phonology of Vowel/Zero Neutralisation in BHA.....	64
3.1 Introduction	64
3.2 Vowel Epenthesis in BHA: Background.....	68
3.2.1 The Dialect: BHA	68
3.2.2 Vowel Epenthesis in BHA	69
3.3 Vowel Epenthesis and the Phonology of Vowel/Zero Neutralisation.....	73
3.3.1 Neutralisation through Vowel Epenthesis	73
3.3.2 Stress and Vowel Epenthesis	76
3.3.3 High Vowel Deletion (HVD)/Low Vowel Raising (LVR) and Epenthesis.....	80
3.3.4 Summary.....	83
3.4 Conclusion.....	83

4	The Phonetics of Vowel/Zero Neutralisation in BHA	84
4.1	Introduction	84
4.2	The Completeness Question.....	86
4.2.1	A Production Experiment.....	86
4.2.2	A Perception Experiment.....	117
4.2.3	General Discussion.....	133
4.3	The Genuineness Question	135
4.3.1	Experimental Artefactuality	136
4.3.2	Statistical Artefactuality	169
4.4	Conclusion.....	181
5	Characterising and Quantifying the Phonetics of Neutralisation.....	183
5.1	Introduction	183
5.2	Characterising the Phonetics of Neutralisation: Labelling Criteria.....	185
5.2.1	Overview	185
5.2.2	Statistical Significance versus Practical Significance.....	187
5.3	Quantifying the Phonetics of Neutralisation: Absolute and Relative Measures 203	
5.3.1	Measuring Central Tendency and Dispersion	203
5.3.2	Towards an Intuitive Notion of Variability Fields.....	212
5.4	Conclusion.....	221
6	A Sketch of the Variability-Field Approach.....	222
6.1	Introduction	222
6.2	Dealing with Phonetic Variability: The Confusion	225
6.3	Phonetic Variability as Fields: The Variability Field Model	231
6.3.1	Overview	231
6.3.2	Structure in Variability	235
6.3.3	Field-Forming Variability	243
6.3.4	Background-Forming Variability	254
6.4	Conclusion.....	259
7	Variability Fields and Vowel/Zero Neutralisation in BHA	260
7.1	Introduction	260
7.2	Fixing Interval Width.....	262
7.2.1	The jnd as a Criterion: Overview	262
7.2.2	The jnd as a Criterion: Figures for Parameters.....	264
7.2.3	A jnd-Based Binning Algorithm	267

7.3	VFM-Based Analysis of Vowel/Zero Neutralisation in BHA	270
7.3.1	Purpose	270
7.3.2	Materials.....	271
7.3.3	The Paradigm.....	272
7.3.4	Results	272
7.3.5	Discussion.....	284
7.4	Conclusion.....	288
8	Summary and Conclusions	289
	APPENDIX A	296
	APPENDIX B.....	298
	APPENDIX C.....	299
	APPENDIX D	301
	APPENDIX E.....	306
	APPENDIX F.....	307
	APPENDIX G.....	308
	APPENDIX H	310
	APPENDIX I	312
	APPENDIX J	314
	REFERENCES	316

List of Tables

Table 2-1: Durational differences between ‘voiced’ and ‘devoiced’ obstruents as reported in a sample of neutralisation studies in the literature	34
Table 2-2: Durational differences between ‘voiced’ and ‘devoiced’ obstruents as reported in a sample of neutralisation studies in the literature [continued]	35
Table 2-3: Summary statistics for the durational differences that reached statistical significance in studies on final devoicing	36
Table 2-4: Existing approaches to the complete-incomplete distinction and qualitative variability as relevant to the phonetics of neutralisation	39
Table 2-5: Mean duration and (SD) values in ms of second pre-tonic and first pre-tonic vowels in Russian (figures based on Barnes 2006: 55)	45
Table 3-1: Consonant inventory of BHA	68
Table 3-2: Vowel inventory of BHA	68
Table 3-3: Summary of conditions determining the quality of the vowel ([V ₂]) inserted to break up word-final consonant clusters in words of the shape /...VCR#/. Conditions include (1) a vocalic environment (i.e., the quality of the lexical vowel /V ₁ /) and (2) a consonantal environment involving different types of word-final sonorants and pre-final obstruents. Consonantal environment is only relevant when the lexical vowel is /a/	73
Table 4-1: The minimal-pair stimulus set in the production experiment	88
Table 4-2: Mean and (SD) values of the five acoustic parameters of the study for both ‘a’ and ‘i’ across literates and illiterates (i.e., all speakers), only literates, and only illiterates	97
Table 4-3: Summary statistics of the paired data of the study including central tendency measures (\bar{X}_{PD} : mean paired difference; $R\bar{X}_{PD}$: relative mean paired difference), variability (SD_{PD} : standard deviation of mean paired difference; SD_{PD} : relative standard deviation of mean paired difference), and effect size (Cohen’s d)	103
Table 4-4: i-F2 mean paired differences (in Hz) for each speaker and as a group average in absolute values	107
Table 4-5: Results of a Wilcoxon signed ranks test for the acoustic parameters of the study for both ‘a’ and ‘i’	108
Table 4-6: Mixed Anovas main effects of V ₂ Underlying Status and Literacy and the interactions between these variables for both ‘a’ and ‘i’ along the five acoustic parameters of the study	111
Table 4-7: Results of two-tailed paired t-tests for ‘a’ and ‘i’ data along the acoustic parameters of the study	112
Table 4-8: The four logically possible scenarios for the phonetics-phonology relation according to neutralisation effects involving [a]-epenthesis and [i]-epenthesis	115
Table 4-9: Speaker L-E’s mean and (SD) values of the acoustic parameters of the study for ‘a’ and ‘i’ by V ₂ Underlying Status; total number of tokens for each parameter= 84 (14 words X 2 V ₂ Underlying Status X 3 repetitions)	119
Table 4-10: Percent correct identification and discrimination of ‘a’ and ‘i’ test words	121
Table 4-11: Mean Rationalised Arcsine Units (RAUs) of the identification and discrimination data for both ‘a’ and ‘i’ according to test alternatives and correct responses	125
Table 4-12: Stimulus-response combinations according to Signal Detection Theory as defined for the identification and discrimination tests in this study	128
Table 4-13: Summary of results for the phonology, acoustics, and perceptibility of the BHA neutralisation data analysed in the study	135
Table 4-14: Survey of production experiments classified by parameters of stimuli and orthography	138

Table 4-15: Experimental variables in the paradigm.....	139
Table 4-16: Experimental conditions in blocks based on stimulus list.....	143
Table 4-17: Stimulus and task effects on acoustic measures for ‘a’ and ‘i’ based on the data summarised in the box-plot graphs below.....	155
Table 4-18: Descriptive statistics for Turkish final devoicing including mean paired differences and their standard deviations as calculated using a by-subject aggregation method (i) and a pooling method (iii); data come from Kopkallı (1993).	175
Table 4-19: Results of two-tailed paired t-tests for Turkish final devoicing.....	176
Table 4-20: Descriptive statistics for Warsaw-Polish final devoicing including mean paired differences and their standard deviations as calculated using a by-subject aggregation method (i) and a pooling method (v); data come from Tieszen (1997).	178
Table 4-21: Descriptive statistics for Kraków-Polish final devoicing including mean paired differences and their standard deviations as calculated using a by-subject aggregation method (i) and a pooling method (v); data come from Tieszen (1997).	178
Table 4-22: Results of two-tailed paired t-tests for Warsaw-Polish final devoicing.....	179
Table 4-23: Results of two-tailed paired t-tests for Kraków-Polish final devoicing	179
Table 4-24: Results of Repeated-Measures Anovas for Warsaw-Polish final devoicing	180
Table 4-25: Results of Repeated-Measures Anovas for Kraków-Polish final devoicing	180
Table 5-1: Statistical significance and sample size: independent-samples t-test (based on Marsh et al 2008)	191
Table 5-2: Statistical significance and sample size: two-tailed paired-sample t-tests	191
Table 5-3: Sample size and percentage of studies yielding statistical significance out of 5000 simulated experiments in each size-condition: Repeated-Measures Design	192
Table 5-4: Mean, SD, C_v , and range values of the four fictitious datasets in Figure 5-1	210
Table 5-5: Mean, SD, C_v , and range values of the four datasets Figure 5-3.....	213
Table 5-6: VFI values of the three hypothetical datasets in Figure 5-6 with and without correcting for bimodality.....	219
Table 5-7: Variability measures applied to the four fictitious datasets histogrammed in Figure 5-1.....	220
Table 5-8: Variability measures applied to the four fictitious datasets histogrammed in Figure 5-3.....	220
Table 7-1: Selected jnd figures for each of the acoustic parameters of the study.....	265
Table 7-2: Summary statistics of i-duration data (n=360) including median (Q2), upper and lower quartiles, and maximum and minimum values	269
Table 7-3: Bin width, bin number, and maximum and minimum of i-duration data read off histograms generated using SPSS settings and VFM’s jnd-based algorithm.....	270
Table 7-4: Stimulus set in terms of phone-fields.....	271
Table 7-5: Experimental conditions in blocks based on stimulus list (reproduced from chapter four).....	272
Table 7-6: VFM summary statistics of the BHA neutralisation data along the acoustic parameters of the study	273

List of Figures

Figure 2-1: Phonetics accessing the input to the phonology (reproduced from Gerfen & Hall 2001: 31).....	58
Figure 2-2: The dynamical models of two grammars; (x) is a continuous articulatory variable (Gafos 2003a: 8).....	59
Figure 2-3: Cooperation and competition between grammar dynamics and intention dynamics in the production of final ‘voiced’ and ‘voiceless’ obstruents in German (reproduced from Gafos 2003a: 16-17).....	61
Figure 4-1: Stimulus pairs defined by the frequency of their members.....	90
Figure 4-2: Mean values of epenthetic [a] vs lexical /a/ along the five acoustic parameters of the study by individual speakers on the left, and over all speakers on the right $\pm 1SD$. For more clarity, the charts on the left do not display SD values.	99
Figure 4-3: Mean values of epenthetic [i] vs lexical /i/ along the five acoustic parameters of the study by individual speakers on the left, and over all speakers on the right $\pm 1SD$. For more clarity, the charts on the left do not display SD values.	100
Figure 4-4: Cohen’s d and 95% CI values for ‘a’ and ‘i’ along the acoustic parameters of the study.....	103
Figure 4-5: Tukey’s mean-difference plots for F0, intensity, and duration of ‘a’ and ‘i’ data (stars represent group values): Difference= $\bar{X}_{epe} - \bar{X}_{lex}$; Mean= $(\bar{X}_{epe} + \bar{X}_{lex})/2$	105
Figure 4-6: Tukey’s mean-difference plots for F1 and F2 of ‘a’ and ‘i’ data (stars represent group values): Difference= $\bar{X}_{epe} - \bar{X}_{lex}$; Mean= $(\bar{X}_{epe} + \bar{X}_{lex})/2$	106
Figure 4-7: Item departures from the group mean values of the acoustic parameters of the study for both ‘a’ and ‘i’	109
Figure 4-8: Number of acoustic parameters with Cohen’s $d > .3 $ for both ‘a’ and ‘i’ by speaker.....	119
Figure 4-9: Identification of ‘a’ and ‘i’ test words according to V ₂ Underlying Status (top) and according to response type (bottom); error bars show 1SD	122
Figure 4-10: Discrimination of ‘a’ and ‘i’ test words according to V ₂ Underlying Status (top) and according to response type (bottom); error bars show 1SD	123
Figure 4-11: Mean Rationalised Arcsine Units (RAUs) of the identification data for both ‘a’ and ‘i’ according to test alternatives and correct responses; error bars show 1SD...125	125
Figure 4-12: Mean Rationalised Arcsine Units (RAUs) of the discrimination data for both ‘a’ and ‘i’ according to test alternatives and correct responses; error bars show 1SD...126	126
Figure 4-13: Mean Rationalised Arcsine Units (RAUs) of the identification (calculated over noun responses) and discrimination (calculated over same responses) data for both ‘a’ and ‘i’; error bars show $\pm 1SD$	126
Figure 4-14: Mean Rationalised Arcsine Units (RAUs) of the identification and discrimination data (calculated over correct responses) for both ‘a’ and ‘i’; error bars show $\pm 1SD$	126
Figure 4-15: Mean A’ for ‘a’ and ‘i’ by perception task; error bars show 1SD.....	128
Figure 4-16: Bias for ‘a’ and ‘i’ by perception task; error bars show 1SD.....	128
Figure 4-17: Mean F0 by experimental condition for ‘a’ and ‘i’.....	147
Figure 4-18: F0 SD by experimental condition for ‘a’ and ‘i’	147
Figure 4-19: Mean intensity by experimental condition for ‘a’ and ‘i’.....	148
Figure 4-20: Intensity SD by experimental condition for ‘a’ and ‘i’	148
Figure 4-21: Mean duration by experimental conditions for ‘a’ and ‘i’.....	149
Figure 4-22: Duration SD by experimental condition for ‘a’ and ‘i’	149
Figure 4-23: Mean F1 by experimental condition for ‘a’ and ‘i’.....	150
Figure 4-24: F1 SD by experimental condition for ‘a’ and ‘i’	150
Figure 4-25: Mean F2 by experimental condition for ‘a’ and ‘i’.....	151

Figure 4-26: F2 SD by experimental condition for ‘a’ and ‘i’	151
Figure 4-27: F0 mean paired differences \bar{X}_{PD} and SD_{PD} values (epenthetic – lexical) according to experimental conditions for ‘a’ and ‘i’	152
Figure 4-28: Intensity mean paired differences \bar{X}_{PD} and SD_{PD} values (epenthetic – lexical) according to experimental conditions for ‘a’ and ‘i’	153
Figure 4-29: Duration mean paired differences \bar{X}_{PD} and SD_{PD} values (epenthetic – lexical) according to experimental conditions for ‘a’ and ‘i’	153
Figure 4-30: F1 mean paired differences \bar{X}_{PD} and SD_{PD} values (epenthetic – lexical) according to experimental conditions for ‘a’ and ‘i’	154
Figure 4-31: F2 mean paired differences \bar{X}_{PD} and SD_{PD} values (epenthetic – lexical) according to experimental conditions for ‘a’ and ‘i’	154
Figure 4-32: Box-plots for a-F0 and i-F0 according to V_2 Underlying Status by experimental condition; the graphs show the median, upper and lower quartiles, and range of ‘a’ and ‘i’ data.....	156
Figure 4-33: Box-plots for a-intensity and i-intensity according to V_2 Underlying Status by experimental condition; the graphs show the median, upper and lower quartiles, and range of ‘a’ and ‘i’ data.	157
Figure 4-34: Box-plots for a-duration and i-duration according to V_2 Underlying Status by experimental condition; the graphs show the median, upper and lower quartiles, and range of ‘a’ and ‘i’ data.	158
Figure 4-35: Box-plots for a-F1 and i-F1 according to V_2 Underlying Status by experimental condition; the graphs show the median, upper and lower quartiles, and range of ‘a’ and ‘i’ data.....	159
Figure 4-36: Box-plots for a-F2 and i-F2 according to V_2 Underlying Status by experimental condition; the graphs show the median, upper and lower quartiles, and range of ‘a’ and ‘i’ data.....	160
Figure 5-1: Histograms of four fictitious datasets with the same distributional shape.....	210
Figure 5-2: Variability in the four datasets in Table 5-4 as measured using SD, C_V , and range values	211
Figure 5-3: Four datasets with the same mean and SD values (adapted from Pingel 1993: 71).....	213
Figure 5-4: Line graphs with the same spread but different gravitation bars.....	216
Figure 5-5: A pyramid graph illustrating the inverse relationship between frequency of the modal interval (shown as a vertical line in the middle of the pyramid) and perceived variability (indicated by shaded bands of different sizes); shading indicates the force of gravitation.	217
Figure 5-6: Histograms of three hypothetical datasets having the same number of intervals but differing with respect to their modal intervals	219
Figure 6-1: Variability as viewed in the literature: Lawful = Non-random.....	228
Figure 6-2: Phonetic Variability according to VFM: Lawful > Non-random.....	231
Figure 6-3: A schematic graphing the constraining effect of a three-way laryngeal contrast on the amount of VOT variation in the realisation of the contrasting sounds (reproduced from Vaux & Samuels 2005: 411)	232
Figure 6-4: Overview of linguistic, para-linguistic, and extra-linguistic effects modelled in terms of VFM.....	234
Figure 6-5: Lawful variability in VFM as forming fields or backgrounds, or as exerting attraction forces on the realisation of the material within a phone-field	235
Figure 6-6: A hypothetical set of normally distributed duration data (n=50) arranged horizontally from smallest to largest (top) and plotted as a histogram with a bin size of 20ms (bottom).....	237
Figure 6-7: Schematics illustrating the bi-directional relationship between composite and primary variability; as indicated by the arrows, adding a source of variation to a set representing primary variability, as in (a), can create composite variability, as in (b-	

d); removing the contribution of a source of variation from composite variability can result in primary variability; distributional properties of a set of variability such as modality (b), shape (c), and spread (d) can be affected by the addition or removing of a source.....	239
Figure 6-8: Histograms illustrating six possible modes of analysis of a hypothetical F2 dataset collected from a single speaker repeating 50 CV items 27 times; Vs= [i], [e], [a], [o], and [u]; Cs= [b], [d], [g], [v], [z], [ʃ], [m], [n], [r], and [w]......	241
Figure 6-9: Histograms illustrating three possible analyses of [ma] data (n=27) plotted in	242
Figure 6-10: Phone-fields in a hypothetical three-word lexicon.....	245
Figure 6-11: A flow chart showing the components of a phone-field.....	246
Figure 6-12: A simple co-ordinate system with three reference points and six different points lying at various distances from the reference points	247
Figure 6-13: Phoneme-fields of a hypothetical three-word lexicon; phone-fields appear as small ovals inside a phoneme-field; meta-phones appear as triangles.....	250
Figure 6-14: A schematic of a word-finally neutralised /t/-/d/ contrast.....	253
Figure 6-15: A schematic of vowel-reduction patterns in four hypothetical languages (LA-LD) according to VFM.....	254
Figure 6-16: Four schematics of the binary variable 'gender' as background.....	255
Figure 6-17: A schematic of the continuous variable 'speaking rate' as background.....	255
Figure 6-18: Autonomous layered-variability fields drawn against speaking rate as background; each layer has its own modal interval and spread already defined; at any point in time, only one layer is activated. The activated layer appears in black.	256
Figure 7-1: Histograms of i-duration data (n=360) generated using SPSS default settings (histogram i) and VFM settings (histogram ii); frequency scale is the same in both histograms.....	269
Figure 7-2: Line charts displaying modality (grey horizontal bars) and range (black horizontal bars) of the BHA neutralisation data along the five acoustic parameters of the study	274
Figure 7-3: Bar chart displaying VFI values of the five acoustic parameters of the study for both 'a' and 'i' data by V ₂ Underlying Status.....	275
Figure 7-4: Bar chart displaying VFI values as percentages for both 'a' and 'i' data by V ₂ Underlying Status.....	275
Figure 7-5: Bar chart displaying VFI values of the five acoustic parameters of the study for both 'a' and 'i' data across V ₂ Underlying Status.....	276
Figure 7-6: Pie chart displaying proportions of VFI values of the five acoustic parameters of the study for both 'a' and 'i' data across V ₂ Underlying Status.....	276
Figure 7-7: Midpoint of modal intervals of a-F0 and i-F0 data according to V ₂ Underlying Status in the six experimental conditions of the study	277
Figure 7-8: Midpoint of modal intervals of a-intensity and i-intensity data according to V ₂ Underlying Status in the six experimental conditions of the study	277
Figure 7-9: Midpoint of modal intervals of a-duration and i-duration data according to V ₂ Underlying Status in the six experimental conditions of the study	278
Figure 7-10: Midpoint of modal intervals of a-F1 and i-F1 data according to V ₂ Underlying Status in the six experimental conditions of the study	278
Figure 7-11: Midpoint of modal intervals of a-F2 and i-F2 data according to V ₂ Underlying Status in the six experimental conditions of the study	279
Figure 7-12: Modal differences in jnd units between epenthetic and lexical vowels in the BHA neutralisation data in the six experimental conditions of the study.....	280
Figure 7-13: F0 and intensity modal-departures from the modal interval of the phone-field by epenthetic and lexical vowels in the six experimental conditions of the study; zero-difference lines are shown as horizontal continuous lines.....	281

- Figure 7-14: Duration modal-departures from the modal interval of the phone-field by epenthetic and lexical vowels in the six experimental conditions of the study; zero-difference lines are shown as horizontal continuous lines.282
- Figure 7-15: F1 modal departures from the modal interval of the phone-field by epenthetic and lexical vowels in the six experimental conditions of the study; zero-difference lines are shown as horizontal continuous lines.282
- Figure 7-16: F2 modal-departures from the modal interval of the phone-field by epenthetic and lexical vowels in the six experimental conditions of the study; zero-difference lines are shown as horizontal continuous lines.283

1 Introduction

1.1 Prelude

On January the 28th, 1986, the space shuttle Challenger exploded just moments after lift-off, killing its crew of seven astronauts. In his analysis of this fatal event, Edward Tufte (1997) made the claim that the shuttle disaster could have been avoided had the data that were used for deciding to launch the rocket on that cold winter day been analysed differently. This is an extreme case where proper data analysis and presentation could have made the difference between life and death. More generally, however, Tufte's claim highlights the crucial contribution of good data analysis and presentation to science and life. Different tools and techniques of data analysis and presentation can reveal as well as obscure important evidence that can guide and improve our understanding of the world.

In this thesis, I take Tufte's point to heart and apply it to a (rather less life-and-death) domain: the phonetics of neutralisation.

1.2 Towards a Better Understanding of the Phonetics of Neutralisation

A central argument of this thesis is that proper data analysis is essential for an adequate understanding of the phonetics of neutralisation. I illustrate this point by showing how different analyses of the same neutralisation data can yield qualitatively quite different results. I use for this purpose (1) an original set of neutralisation data from a Bedouin Hijazi dialect of Arabic (henceforth BHA) and (2) datasets reported in the literature from Turkish and Polish.

With reference to the Turkish and Polish datasets, I demonstrate the effect of pre-analysis on the statistical outcomes of analyses conducted using the same inference tool (statistical significance tests), and subsequently on the inferences we draw from such analyses. I show that, by adopting different data-entry procedures, we arrive at interpretations of the phonetics of neutralisation for the Turkish and Polish data that are distinct from what is reported in the literature.

Kopkallı (1993) found complete neutralisation for final devoicing in Turkish. That is, final voiceless and devoiced stops are not statistically significantly different along the durational dimensions that Kopkallı measured for her data. Kopkallı's conclusion is based on separate analyses of data produced by each individual speaker in the study. In this thesis, I re-analyse the same data but reach a completely different conclusion. On the basis of statistical tests run on by-subject aggregated data as well as on pooled data from the participants in the Kopkallı study, I conclude that neutralisation in Turkish is phonetically incomplete. That is, there are small but statistically significant differences between word-final voiceless and devoiced stops in Turkish.

The reverse scenario obtains for the neutralisation data from Warsaw Polish as reported in Tieszen (1997). Tieszen reported incomplete neutralisation on the basis of the outcome of statistical tests run on by-item aggregated data. I re-analyse the same dataset as a by-subject aggregate but find complete neutralisation.

With regard to the neutralisation data from BHA, I take the argument beyond the treatment of pre-analysis data and straight to the heart of the inference mechanism itself, including its descriptive quantification and inference procedures and format. I consider two tools: (1) Null Hypothesis Significance Testing (henceforth NHST), which is the standard evidence-finding procedure in phonetic research, and (2) the Variability Field Model (henceforth VFM), which is a new model of phonetic processing and representation I propose in this thesis. VFM makes no reference to statistical significance. It introduces a novel phonetics-specific quantification of phonetic data and variability. According to VFM, phonetic data are examined as intervals rather than as single data points. To divide a dataset into intervals, VFM applies a binning algorithm that utilises the familiar psycho-

physical notion of just noticeable difference (jnd). An immediate advantage of this move for phonetic data analysis is that differences that are not even just-noticeable can no longer clutter our analysis and presentation of data, thus allowing structure and patterns present in a given dataset to emerge more distinctly. A central VFM argument is that to find structure in variability, we need only consider hearable variability. Moreover, VFM places emphasis on count data rather than arithmetically derived data like the mean and the standard deviation (SD), which standardly summarise phonetic data within NHST. This manoeuvre on the part of VFM brings the model more in line with frequency-based Bayesian reasoning (Gigerenzer & Hoffrage 1995).

To appreciate how these analysis tools can lead us to incompatible conclusions regarding the phonetics of neutralisation in BHA, it is instructive to familiarise ourselves with the neutralising effects of vowel/zero alternations in BHA, which constitute the empirical focus in this thesis.

In BHA, certain underlying consonant clusters are broken up through the epenthesis of a vowel of predictable quality, e.g. /laħm/ > [laħam] 'meat', /gidr/ > [gidir] 'pot'. The epenthesised vowels potentially neutralise with lexical vowels. For example, [laħam] can be the output of either /laħm/ 'meat' (with epenthetic [a]) or /laħam/ 'shut tight'; [gidir] can be the output of either /gidr/ 'pot' (with epenthetic [i]) or /gadir/ 'managed/overpowered' (with independent raising of the first vowel). It is important to insert the caveat 'potentially' because BHA phonology treats epenthetic vowels differently according to their quality. For example, while both epenthetic [a] and lexical /a/ are stressable, epenthetic [i], unlike lexical /i/, is not.

NHST-generated results reveal a curious pattern of dis-correlation between the phonetics and phonology of neutralisation in BHA. The tests find a statistically significant acoustic difference where it is least expected phonologically, while they fail to find statistically significant acoustic differences between the vowels that are phonologically different. Specifically, by the criterion of statistical significance, epenthetic [a] and lexical /a/, which behave the same in the phonology, are distinct in the phonetics. Conversely, epenthetic [i] and lexical /i/, which are treated

differentially in the phonology, are acoustically non-distinguishable by that criterion of statistical significance.

By way of contrast, according to the VFM results, the neutralisation effect through [a]-epenthesis, which is phonologically complete, is also phonetically complete. Conversely, the underlying distinction between epenthetic [i] and lexical /i/, which survives in the phonology, also survives in the phonetics. In other words, the VFM analysis yields a good correlation between the phonetics and phonology of vowel/zero neutralisation in BHA.

Which of these analyses is right? Of course the answer to this question is never going to be as critical as that given to the questions about the data analysis and decisions in the Challenger catastrophe in 1986. Yet, an adequate understanding of neutralisation data and analysis tools is necessary for a better understanding of the phonetics of neutralisation, which this thesis seeks to attain.

The purpose of this thesis is to provide a better-informed qualitative and quantitative description of the phonetics of neutralisation. To that end, the thesis reaches for insights from research fields as diverse as statistics, cognition, neurology, and psychophysics, in addition, of course, to phonetics and phonology. Much closer to home, though, a deep acquaintance with the empirical literature on the phonetics of neutralisation has not only guided the development of the thesis but has also revealed many of the issues that I investigate here.

These issues concern the extent of phonetic merger, the empirical validity of the phonetic effect, and the variability inherent in neutralisation data. The first two of these have attracted a lot of attention in the literature. I refer to these two issues respectively as the completeness question and the genuineness question. One contribution of this thesis involves addressing neglected aspects of these questions. By doing so, the thesis fills a gap in the literature and raises, at the same time, new concerns over how we approach, describe, and reach conclusions on the phonetics of neutralisation.

Firstly, regarding the completeness question, I focus on the relation between the phonetics and phonology of neutralisation. I investigate the phonetics of two neutralisation scenarios bearing directly on the completeness question, with a view to bringing more balance to our approach to the relation between the

phonetics and phonology of neutralisation. As a schematic description, scenario A, which has attracted most experimentation, involves contrasts that are impressionistically described as neutralised, with surface phonology completely giving up on them. Conversely, scenario B, which is under-researched from an experimental perspective, involves contrasts which are impressionistically described as neutralised, but which the surface phonology still refers to under certain conditions. One important condition is their role in the creation of what is known in the phonological literature as underapplication or overapplication opacity (e.g., Kiparsky 1973; McCarthy 1999). Until now, these scenarios have been investigated separately. But here I look more closely at both, using data from a single phonological process drawn from a single language (see §1.3.2 for details).

Secondly, in connection with the genuineness question, I engage in the debate from an experimental and statistical perspective. Experimentally, I devise a paradigm that manipulates important variables claimed to influence the phonetics of neutralisation. These variables include orthography, pragmatic context, and the presence of minimal pairs in the stimulus list. The main conclusion I reach is that the same experimental make-up can produce both complete and incomplete effects, with no definitive correlation between either effect and experimental artefactuality.

My re-analysis of the neutralisation data reported in the literature from Turkish and Polish provides the statistical side of my examination of the genuineness question. My main conclusion on this issue is that arguments questioning the genuineness of the reported findings are self-defeating, irrespective of whether they appeal to experimental or statistical considerations.

Thirdly, the thesis investigates the issue of variability, which, despite its obvious relevance to neutralisation data, has only received the slightest attention in the relevant literature. I note here that variations due to speakers, items, and conditions are systematically reported, but that their implications for the phonetics of neutralisation are no less systematically neglected.

The thesis discusses the issue of variability over the space of three main chapters. The approach underlying VFM, as described above, calls for the integration of

variability into any adequate model of the phonetics of neutralisation. VFM treats variability as the essence of phonetic data rather than some isolatable addition.

I suggest that we can make far better use of the experimental findings in the neutralisation literature by paying more attention to the phonetic variability they contain. This, I argue, will place us in a better position to answer basic questions about the extent of phonetic merger and the genuineness or otherwise of (in)complete neutralisation and to evaluate competing proposals for modelling the phonetics of neutralisation. Also, and perhaps more importantly, it will allow us to appreciate better what the phonetics of neutralisation can tell us about the relation between phonetics and phonology and between speech production and perception and about the place of these relations in the mental lexicon.

In the remainder of this introductory chapter, I present glimpses of the main features of the laboratory tradition in studies of neutralisation. I give a background introduction in §1.3.1 and move on to discuss the phonetics-phonology relation as depicted in the literature on the phonetics of neutralisation in §1.3.2. The section also previews the phonology of vowel/zero neutralisation in BHA and points out its empirical relevance to the phonetics-phonology relation. In §1.4, I close by laying out the structure of the thesis.

1.3 Neutralisation in Laboratory Studies

1.3.1 Background

Among the issues that a laboratory study of neutralisation should aspire to elucidate are the phonetics-phonology relation, the production-perception relation, and the relation between phonetic variability and lexical representation and learnability. The past forty years or so have produced more than eighty studies¹ subjecting neutralisation to experimentation, but the discussion of most of the above-mentioned issues remains largely unformed by the findings of these studies. A factor immediately contributing to this unfortunate state of affairs is that these questions were simply not on the research agenda for many of those who set

¹ The vast majority of published papers are found in the *Journal of Phonetics*, *Phonetica*, the *Journal of the Acoustical Society of America*, and *Research in Phonetics*.

out to study neutralisation from an experimental perspective. Generally, the purpose of most laboratory inquiries into neutralisation was to test what was treated as the null hypothesis in (1):

- (1) Neutralisation is phonetically complete in that no phonetic traces of the relevant underlying contrast are to be found when inspecting the acoustic² signals of the terms of the neutralised contrast (along the acoustic parameters to be investigated).

This all followed the discovery of incomplete neutralisation, the very instantiation of the alternative hypothesis to (1). Consider the following excerpt from the opening paragraph of a paper by Fourakis and Iverson (1984) entitled “On the ‘incomplete neutralisation’ of German final obstruents”:

Until very recently, there was no reason for the present paper to have been written, the simple purpose of which is to establish whether the devoicing of word-final obstruents in German constitutes phonological neutralization. This just has not been a serious question in modern linguistics (p. 140).

However, the genuineness of incomplete neutralisation was soon questioned on the grounds that the acoustic residue was found to fluctuate to the point of disappearance under different experimental conditions (see e.g., Dmitrieva et al 2010; Jassem & Richter 1989; Port & Crawford 1989; Snoeren et al 2006; Warner et al 2006, 2004). Blame for this mostly fell on the specifics of the experimental design which, according to some, encouraged ‘spelling pronunciation’ (e.g., Fourakis & Iverson 1984; Warner et al 2006) and/or ‘hypercorrection’ (e.g., Jassem & Richter 1989). Both of these were thought to issue from what Barnes (2006: 225) called the ‘paralinguistic contamination problem’.

Another, far less celebrated, source of scepticism concerning the genuineness of incomplete neutralisation is that the reported phonetic differences are not always in the same direction as when the relevant contrast is fully realised (e.g., Barnes 2006; Dinnsen & Charles-Luce 1984; Wissing & van Rooy 1992, in van Rooy et al 2003). These inconsistencies added to the controversy that was already engulfing incomplete neutralisation, especially owing to the fact that a number of early

² Articulatory differences have also been reported in the literature (see e.g., Butcher 1995; Nolan 1992; Smorodinsky 2002). However, for the sake of brevity, I only discuss acoustic studies here.

studies yielded mixed results even when they were investigating the same neutralisation case (see e.g., Fourakis & Iverson 1984 vs Port & O'Dell 1985 for *German*; Jassem & Richter 1989 vs Slowiaczek & Dinnsen 1985 for *Polish*; Baumann 1995 vs Warner et al 2004 for *Dutch*; Port 1977 vs Fox & Terbeek 1977 for *American English*).

After decades of experimentation, the basic exploratory questions of the reality and mechanism of incomplete neutralisation are far from settled. Commentators on the topic continue to call for further research. For example, Barnes' 2006 book on positional neutralisation contains a discussion of incomplete neutralisation as a challenge to any model of the phonetics-phonology interface. He describes "studies seeking to dispel once and for all persistent scepticism concerning the paralinguistic contamination problem for I[ncomplete] N[eutralisation]" as "extremely important" (p. 234). Similarly, Ernestus (in press: 10) deems "further research into this issue [...] indispensable".

1.3.2 Neutralisation and the Phonetics-Phonology Relation

As I pointed out above, the most significant outcome of decades of phonetic inquiry into neutralisation is that it may be phonetically incomplete, with traces of the underlying contrast surviving in the acoustic signal. To many phoneticians, the discovery of incomplete neutralisation has strengthened calls for a greater role for phonetics in phonology. For some time it looked as though, upon the recognition of the existence of incomplete neutralisation, "[a] theoretical dilemma [for phonology...] arises concerning the presumed distinction between neutralization rules and allophonic rules", which only phonetics can resolve (Slowiaczek & Dinnsen 1985: 326) (but see Blumstein 1991 and Nolan 1986 for a different view). This justified the investment of time and effort in small-scale studies whose main mission was to prove or disprove the existence of incomplete neutralisation. Discussions of the implications of these findings for the relationship between phonetics and phonology have had little impact. Furthermore, the credibility of the studies was considerably undermined by the identification of a number of methodological shortcomings, with the potential to call into question the genuineness of the reported findings (see §1.3.1 for references).

To many phonologists, on the other hand, the discovery of incomplete neutralisation is only very recently beginning to assume any theoretical relevance. Incomplete neutralisation is now being taken to provide support to a number of phonological models including the Dynamics model (Gafos 2006; Gafos & Benuš 2006; Nycz 2005), Candidate Chains in Optimality Theory (Gouskova & Hall 2009), Turbidity Theory (van Oostendorp 2008), phonologisation of sound change in progress (e.g., Barnes 2006), and the recent application of exemplar-based models to neutralisation data (e.g., Ernestus & Baayen 2006; cf. Yu 2007).

But before that, and apart from a total rejection by phonologists like Manaster Ramer (1996a, 1996b), findings of those instrumental studies on neutralisation received little attention from the phonological community. Dinnsen (1985) attributes this undue disregard on the part of phonologists to their insistence on categoricity, which leads them, he claims, to assign priority to perception over production. Dinnsen protests that, for contemporary phonologists, “if for some reason production differences were discovered without the perception differences, those differences would be dismissed as linguistically irrelevant” (p. 273).

On the other hand, Brockhaus (1995) observes that at the time when neutralisation was becoming empirically relevant to many phoneticians in the 80’s and 90’s, its theoretical relevance to phonologists was declining.

Had the purpose of laboratory inquiries into neutralisation been a quest for insights from production and perception data into covert and ‘problematic’ phonological patterns, the “prospect of having to do phonology or phonetics in a world in which ‘incomplete neutralization’ really existed” might have been less “horrific” to phonologists like Manaster Ramer (1996b: 514). Unfortunately, laboratory studies of neutralisation fail to deliver on this objective. Their approach to the production-perception relation is biased (as pointed out by Dinnsen 1985); so is their approach to the phonetics-phonology relation. As mentioned above, the null hypothesis most of these studies are testing is that phonological neutralisation is phonetically complete. What this implies for the phonetics-phonology relation is that the phonetics and phonology of neutralisation mirror each other. More specifically, phonetics is assumed to reflect the phonological surface representation, where neutralisation occurs, and be blind to the underlying representation, where contrast holds. Most researchers choose for instrumental

investigation contrasts that are reportedly neutralised, where phonology completely gives up on the underlying contrast on the surface. To place it within its phonological context, this would be an illustration of clause (c) of Kiparsky's (1973: 79) definition of opacity³ given in (2). Importantly, this clause depicts neutralisation as opacity. Opacity here refers to the problem of learning the correct underlying representation of a contrast that is neutralised on the surface.

- (2) A process P of the form $A \rightarrow B / C_D$ is opaque to the extent that there are phonetic forms [i.e., surface representations] in the language having
- (a) A in the environment C_D [underapplication opacity], or
 - (b) B derived by P in environments other than C_D [overapplication opacity], or
 - (c) B not derived by P in the environment C_D [neutralisation]⁴.

Interestingly, clause (c) was only later added to the original definition that appeared in Kiparsky (1971: 621-622). However, this clause has systematically been ignored by phonologists who discuss opacity. Those who ever mention it hasten to announce its irrelevance to their discussion of the problematic, thus more interesting, cases involving the other two clauses (see e.g., Baković 2007; Bermúdez-Otero 1999; McCarthy 1999). The task of the investigator of the phonetics of a neutralisation effect that conforms to clause (c), then, is to find out if the relevant neutralisation, which is complete in the phonology, is also complete in the phonetics.

³ For expository convenience, I follow the numbering of clauses that appears in McCarthy (1999: 358). See also Baković (2007: 219).

⁴ More concretely, a process deriving, for example, [t] from an underlying /d/ in a certain context is opaque if there exists within the language and in the same context an underlying /t/ surfacing as [t] or a [t] derived by a rule other than the /d/ → [t]. In other words, in a language of this type, *only some* surface [t]s are underlyingly /t/s. Note that this view of neutralisation as involving opacity makes little sense under a broader conception of neutralisation that does not distinguish between 'the obliteration of contrasts that exist at the lexical level' and 'the static lack of contrast at the lexical level'. An important question to ask here is whether or not such a distinction is warranted. Although this question is seldom asked in the phonological literature (but see Gurevich 2004; Hansson 2008), it seems to me that these two 'patterns' make different predictions regarding the phonetics of neutralisation, and thus should be kept as separate in phonetic investigations of neutralisation. More specifically, neutralisation in the narrow sense as *the obliteration of lexical contrasts on the surface* predicts that the realisation of a neutralised contrast may be subject to pressures from a number of factors including, for example, the different URs of the terms of the contrast, the existence of morphologically related words where the underlying contrast is fully maintained on the surface, and the different orthographic forms where the underlying contrast is represented (cf. Guskova & Hall 2009; Yu 2009).

However, it is not hard to see that the correlation between the phonetics and phonology of neutralisation assumed by most researchers could also be tested by studying cases where a phonetically incomplete neutralisation *should be* what we expect if the hypothesis of a close correlation is to be retained. Reviewing the literature on the phonetics of neutralisation, we find that only a few studies have investigated contrasts whose neutralisation plays a role in the creation of underapplication or overapplication opacity (clauses (a)/(b) of the definition in (2) above). These studies focus on a few phonological processes such as vowel epenthesis (e.g., Gouskova & Hall 2009), flapping (e.g., Colley 2009; Fox & Terbeek 1977), and final devoicing (e.g., Baroni & Vanelli 2000; Ettliger 2007). In these cases, if the phonetics and phonology of neutralisation mirrored each other, we would expect traces of the underlying contrast to survive in the phonetics, as they do in the phonology.

The importance of testing the difference between the two takes on the correlation of the phonology and phonetics of neutralisation, by focusing on a single phonological process in a single language, should be obvious. On the one hand, such an investigation would help shape and sharpen our theoretical perspective on the phonetics of neutralisation and more generally the phonetics-phonology relation. On the other, it should be empirically beneficial in the context of the genuineness question mentioned above. Fortunately, just such a testing-ground exists and is the subject of investigation in this thesis.

In Bedouin Hijazi Arabic (BHA), vowel epenthesis neutralises vowel/zero contrasts differently according to the quality of the inserted vowel. The result of epenthesising [a] in a word-final bi-consonantal cluster that violates the Sonority Sequencing Principle (SSP) (Clements 1990: 285) is apparently a total obliteration of the underlying vowel/zero contrast on the surface: an epenthetic [a], underlyingly a zero, and a lexical /a/ are both impressionistically described and transcribed as being the same, and crucially, they behave the same with respect to phonological processes. Or to be more precise, there is no process in the phonology of BHA that treats epenthetic [a] and lexical /a/ differently. Take for example stress assignment. Epenthesising [a] does not disrupt stress assignment. Both epenthetic [a] and lexical /a/ are stressable:

(3)	/laħam+ha/	[laħámha]	'shut it tight'
	/lahm+ha/	[laħámha]	'her meat'

This is not what we observe when the epenthetic is [i]. Epenthesising [i] to break up a bi-consonantal cluster that violates SSP results in the same vowel/zero contrast being neutralised, but not quite: despite sounding the same to the naked ear, lexical /i/ and epenthetic [i] behave differently with respect to some phonological processes. Sticking with stress assignment for an initial illustration, we note that [i]-epenthesis renders stress assignment opaque or non-surface-true in McCarthy's (1999) sense:

(4)	/gadir+ha/	[gidírha]	'overpowered her'
	/gidr+ha/	[gídírha]	'her pot'

Relating these phonological facts to the phonetics of vowel/zero neutralisation in BHA, we come up with the predictions in (5), which the phonology seems to support.

- (5) Assuming that the phonetics and phonology of neutralisation correlate closely, epenthetic [a] and lexical /a/, which behave phonologically the same, are predicted to be acoustically and perceptually non-distinguishable, whereas epenthetic [i] and lexical /i/ are predicted to be acoustically and perceptually different since BHA phonology treats them differently.

Put differently, the case we have for empirical investigation involves a single phonological process neutralising the same type of contrast differently. The results we get by studying this controlled case that BHA phonology offers will bring the phonetics and phonology of neutralisation face to face in a manner unreported before. This will allow us to add a new twist to the old story of the phonetics of neutralisation, as told from the angle of the completeness question, the genuineness question, and the variability question.

1.4 Structure of the Thesis

The thesis proceeds as follows. In chapter two, I introduce the notion of qualitative variability that arises as we approach the phonetics of neutralisation with a predisposition toward interpreting the results in a way that categorically distinguishes between complete and incomplete neutralisation. I then discuss and evaluate the various predictions made by the few models of the phonetics-phonology interface that have attempted to describe the phonetics of neutralisation. My assessment reveals an intolerance of qualitative variability shared by the models I review.

In chapter three, I document the phonology of the relevant vowel/zero alternations in BHA. The chapter opens with a brief description of the dialect, noting its place in phonological work to date. An overview of vowel epenthesis in BHA follows. The main part of the chapter comprises a detailed look into vowel epenthesis, vowel/zero neutralisation, and the phonological interactions that illustrate the disparate behaviour of epenthetic and lexical vowels in BHA phonology.

In chapter four, I deal with both the completeness and the genuineness questions in light of the phonetics of vowel/zero neutralisation in BHA. In connection with the completeness question, I present acoustic and perceptual analyses of the neutralisation data in BHA and discuss the implications of these analyses for the phonetics-phonology relation. As to the genuineness question, I deal with both its experimental and statistical aspects. Experimentally, the chapter analyses acoustic data generated within an experimental paradigm that explores the effect of orthography, pragmatic context, and stimulus composition. Statistically, I discuss the mixed findings that I derive by subjecting datasets from Turkish and Polish to different pre-analysis procedures.

In chapter five, I present a qualitative and quantitative description of the phonetics of neutralisation. On the qualitative side, I extend the notion of qualitative variability introduced in chapter two. I present a brief scrutiny of the labelling criteria in the literature as applied to the phonetics of neutralisation. I then discuss in detail statistical significance, the single most important criterion that has been standardly used to classify the outcomes of quantitative studies of neutralisation.

On the quantitative side, I evaluate the parametric measures of central tendency and dispersion which are commonly used to quantify the phonetics of neutralisation. I show that the standardly used measures of central tendency (i.e., the mean) and variability (i.e., the standard deviation) are unintuitive and closely tied to the numerical values of the measurement scale in such a way that a robust estimation of the underlying central location and variability can sometimes be hard to attain. I also propose an alternative that is both more intuitive and cognitively plausible. Specifically, I suggest that the mode, rather than the arithmetic mean, be used to measure central tendency. This suggestion rests on the claim that the mode reflects better the intuitive notion of average. By the same token, the variability measure I propose relates the frequency of the modal interval to the range, both of which are more consistent with frequency-based Bayesian reasoning (Gigerenzer & Hoffrage 1995). Moreover, I suggest that phonetic data are more appropriately examined as intervals rather than as single points.

In chapter six, I present a sketch of VFM, summarising its main philosophy and describing how it treats the variability effects modelled in the chapter. These include allophonic and indexical variations. I also detail the field-forming and background-forming components of the model and describe the relationships among them.

Chapter seven is basically a preliminary implementation of the ideas outlined in chapters five and six. Specifically, the chapter provides technical details of how to find the mode for phonetic data and how to calculate the variability field index. It also offers a preliminary VFM-based analysis of neutralisation data from BHA.

Finally, in chapter eight, I summarise the main findings and point out directions for future research.

2 Variability in the Phonetics of Neutralisation

2.1 Introduction

A substantial portion of the experimental literature that has begun to accumulate since the 1970s aims to establish the phonetic completeness or incompleteness of various neutralisation effects. An extensive list of references is provided in Appendix A. Generally, a neutralisation effect is said to be phonetically incomplete if phonetic differences between the relevant sounds are found to be statistically significant; otherwise, neutralisation is said to be phonetically complete.

Experimentation has focused on a diverse range of neutralising processes such as *final devoicing* in German, Catalan, Polish, Russian, Dutch, Turkish, Lithuanian, Friulian, and Afrikaans (for a recent review on final devoicing, see Ernestus, in press), *vowel deletion* in French and Serbo-Croatian, *vowel reduction* in Russian, Catalan, and Shimagonde, *consonant deletion* in Turkish, *vowel epenthesis* in Levantine Arabic and Brazilian Portuguese, *stop epenthesis* in English, *assimilation* in French, Catalan, Russian, English, and Lithuanian, *flapping* in American English, *coda neutralisation* in Andalusian Spanish, Puerto Rican Spanish, and Korean, and *tone sandhi* in Mandarin and Cantonese. See Appendix A for references.

With both complete and incomplete neutralisation reported in the literature, a phonetics-based classification of neutralisation effects seems to have emerged—a classification that is to be added to the list of terminological distinctions that the notion of neutralisation has thus far included. However, it should be noted that older distinctions mostly serve a descriptive purpose—to capture adequately typological patterns of asymmetry in neutralisation phenomena. These distinctions are made along dimensions such as the following: global versus positional (e.g.,

Steriade 1999; Warner et al 2006), productive versus unproductive (e.g., Lombardi 1994), obligatory versus optional (e.g., Kirchner 1998), contextual versus structural, assimilatory versus nonassimilatory (e.g., Trubetzkoy 1969; Harris 1997; Steriade 1999), virtual versus actual (e.g., Hansson 2008), and exceptionless versus exception-sensitive (e.g., Kager 2008). To these dimensions, phonetic research has added the extent of phonetic merger along which the complete-incomplete distinction is defined. This dimension continues to be investigated in sociolinguistic work documenting mergers and near-mergers (e.g., Gordon 2002; Hay et al 2006; for a recent review, see Hall-Lew 2009)⁵.

For scope and time reasons, I focus exclusively on the phonetics of neutralisation as reported in phonetic research. However, my view on merger and neutralisation is very similar to Yu's (2007). Basically, phonetic studies of mergers and neutralisation share an interest in the question of the presence or absence of phonetic differences between sounds thought to have been merged or neutralised. Of course, researchers concerned with mergers and neutralisation may employ different criteria in establishing whether a difference is present or not, use different stimulus materials, different labels, etc. Yet it remains true that most of their differences are methodological.

Undoubtedly, as I pointed out in chapter one, the complete-incomplete distinction is significant not only because of its non-impressionistic, case-study methodology, but also because it raises fundamental theoretical and empirical questions. These questions bear on the reality and mechanism of neutralisation. They also implicate the very concept of contrast, which still plays a central role in phonological theory (for more on this, see Hall 2007 and Hall 2009). Moreover, this distinction has considerable potential to yield insights into the relation between phonetics and phonology and between speech production and perception and into the place of these relations in the mental lexicon.

It might seem surprising that a classification with such a substantial claim to both theoretical and empirical legitimacy as the complete-incomplete distinction should

⁵ Sociolinguists have long been interested in the phonetics of mergers and near-mergers (for a review, see Labov 1994). Early reports of near-mergers appear in Labov et al (1972), Trudgill (1974), Milroy and Harris (1980), and Nunberg (1980) among others.

remain a matter of continued dispute (see the next sections for details).⁶ While it is true that the central research question that the bulk of the experimental literature on the phonetics of neutralisation tries to answer presupposes such a classification, at least as a conceptual possibility that needs to be tested, researchers are still divided on what effect to accept as genuine, and what effect to dismiss as artificial, and on what grounds. For example, those who contend that neutralisation may not be phonetically incomplete in real-life situations generally argue that studies claiming otherwise are simply presenting an experimental artefact as a fact (e.g., Manaster Ramer 1996a, 1996b); on the other hand, those to whom neutralisation may not be complete argue that studies reporting complete neutralisation have simply failed to reveal an effect that is present. They explain that complete neutralisation is an artefact of the chosen tools of analysis, phonological or statistical (e.g., Dinnsen 1985; Port 1996; Port & Leary 2005). I discuss the genuineness of complete and incomplete neutralisation from both an experimental and statistical perspective in chapter four.

But, in this chapter, I focus on a different set of ‘curiosities’ which can further chip away at the credibility of the complete-incomplete distinction in the literature.

An impression that deepens the more we read the relevant literature is that the distance separating phonetically complete and incomplete effects can indeed be very small and variable (cf. Warner et al 2004), irrespective of the phonetic dimension along which it is measured. For example, quantitatively, the magnitude of phonetic differences between sounds that are subject to neutralisation as reported in experimental studies varies a lot. Take for instance durational differences in studies investigating the phonetics of final devoicing. Here, reported differences that reach statistical significance range from |1|ms (Charles-Luce & Dinnsen 1987) to |70|ms (Baroni & Vanelli 2000), with |5|ms being the most frequent difference that is statistically significant.⁷ Qualitatively, variability is seen

⁶ Interestingly, a number of researchers have tried to find a link between this phonetics-based distinction and phonology-based distinctions within the notion of neutralisation. For example, Warner et al (2006) hint at the possibility that incomplete neutralisation may be restricted to positional neutralisation, whereas global neutralisation may be phonetically complete. Dinnsen (1985) asserts that non-assimilatory neutralisations are phonetically incomplete. He speculates towards the end of his paper that assimilatory neutralisations (for which he acknowledges he lacks data) might turn out to be complete or incomplete. Similarly, Charles-Luce (1986) wonders whether complete neutralisation is best defined as an assimilatory process in speech production.

⁷ These statistics are given in absolute values.

in terms of the very existence and directionality of statistically significant differences. Some studies report statistically significant differences that have escaped other studies investigating the same neutralisation effect in the same language. The consequence of this is that the same neutralisation case is labelled as phonetically incomplete on the basis of conclusions drawn from studies revealing statistically significant differences, but as phonetically complete on the basis of conclusions drawn from studies reporting no statistically significant differences. For example, Port and O'Dell (1985) conclude that final devoicing in German is phonetically incomplete, whereas Fourakis and Iverson (1984) report that it is phonetically complete. Similarly, Slowiaczek and Dinnsen (1985) conclude that final devoicing in Polish is phonetically incomplete, whereas Jassem and Richter (1989) report that it is phonetically complete. Of course design differences exist among these studies. Whether such differences are wholly responsible for the mixed findings is a question that I attempt in chapter four. Note, however, that the empirical fact of variability remains valid regardless of whether or not we are able to explain it.

As to the directionality of phonetic differences, some studies report differences between neutralised sounds going in the same direction as when the relevant contrast is fully realised. For example, Port and O'Dell (1985) report a statistically significant durational difference in stop closure between 'voiceless' and 'devoiced' final stops, where voiceless stops are on average 5ms longer. In contrast, Dinnsen and Charles-Luce (1984) report a similar statistically significant durational difference but in the opposite direction, with 'voiceless' obstruents being shorter by 7ms than the corresponding devoiced obstruents.

Faced with such inconsistencies, we may find it necessary to dissect experimental design, analysis procedure, and inferential criteria in search of possible explanations. In chapters four and five, I present empirical data which guide a scrutiny of both methodology and inference as found in the phonetic research on neutralisation. Yet, in the face of inconsistencies, we may also find it instructive to explore the variability issue itself. For a start, we need to consider how different approaches taken towards it inform the modelling of the phonetics of neutralisation. This is what I do in this chapter. Specifically, I claim that we can make far better use of the experimental findings in the literature by learning more

about the variability they contain. This variability, I argue, is to be understood not only in a quantitative sense, which is expected and accepted, but also in a qualitative sense. In this thesis, the notion of qualitative variability refers to discrepancies in the conclusions researchers draw on the basis of different statistical analyses. These discrepancies pertain to qualitative aspects of the phonetics of neutralisation, which include the presence and directionality of statistically significant differences. In other words, the qualitative aspects of the phonetics of neutralisation are relevant to the following questions. Is there a statistically significant difference? How prevalent is it? How reliable is its presence? Is the difference in the expected direction? In this sense, qualitative variability is distinct from quantitative variability, which typically concerns quantities in the phonetics of neutralisation. These include the size of the difference and its variance. Relevant questions here include the following. How large is the difference? How variable is it? How much does it vary relative to a reference value like the mean? As we can see, qualitative variability is an ‘acquired’ property of the phonetics of neutralisation that comes about when we abstract away from quantitative variability in our study of the phonetics of neutralisation.

The rest of the chapter proceeds as follows. In §2.2, I present a brief discussion of how qualitative variability is manifested in the phonetics of neutralisation. In §2.3, I offer a detailed and critical look into how the issue of variability has been approached in the literature. I conclude in §2.4.

2.2 Manifestations of Variability in the Phonetics of Neutralisation

As commonly agreed, the phonetics of neutralisation is variable. Now if ‘variable’ is only meant to apply to raw measurement values (e.g., acoustic data), the truthfulness of the claim above can hardly be denied. However, its theoretical or practical relevance can hardly be justified in a study that is meant to inform the debate on the phonetics-phonology relation. Currently, the received wisdom is that acoustic data along any phonetic dimension vary for both linguistic and non-linguistic reasons, both systematically and randomly (see chapter six for details).

Most laboratory studies on the subject aim to confirm or disconfirm systematic variation that can be shown to be attributable to some phonological contrast initially assumed to have neutralised. Determining eventually whether or not the phonetics of a certain neutralisation still bears marks of the purportedly neutralised contrast marks a shift in the focus of these studies from the quantitative domain to the qualitative domain. Now, the output of a long statistical procedure, which is numerical in most situations, is often interpreted in a categorical fashion to mean phonetically complete or incomplete neutralisation. Put differently, a complete-incomplete distinction which is supposed to be statistically demonstrable seems to be made in the phonetics of neutralisation.

However, it should be awkward for this distinction when the phonetics of neutralisation fails to yield conclusive evidence as to the phonetic pattern of the relevant neutralisation, such as when the phonetics of neutralisation continues to propagate mean differences whose numerical values fail to consistently form a single nominal value (i.e., complete or incomplete neutralisation). That is, if studies continue to yield mixed results regarding the overall phonetic pattern of a certain neutralisation, should we blame this inconsistency on methodological and non-linguistic influences, implying that the qualitative outcome of similar studies on the phonetics of neutralisation is essentially replicable? How should we react when findings coming from experimental studies applying the same methodology still fail to converge? How justifiable is it to question instead the reproducibility of the qualitative aspects of the phonetics of neutralisation? It may be the case that the phonetics of neutralisation is essentially variable not only regarding the size of effect, but also, and perhaps more crucially, regarding the existence and direction of statistically significant differences. Let us focus on final devoicing by way of illustration. Table 2-1 and Table 2-2 present durational differences between ‘voiceless’ and ‘devoiced’ obstruents as reported in a sample of studies listed in the tables. The argumentation below makes use of mean difference values, which are available from all the references in the tables, and are sometimes the only type of descriptive statistics given. Needless to say, statistical significance is not solely based on mean differences. However, other things being equal, there is a correlation between statistical significance and the size of mean differences—a correlation that can be defined mathematically, conceptually, and practically. But for the durational differences in the tables below, we do not know if other things

are equal. Summary statistics pertaining to, for example, variation, sample size, and effect size are often not reported (see chapter five for details).

Language	Study	Parameter	Difference (ms)
German	Mitleb (1981)	V-duration (__ stop)	23*
		V-duration (__fricative)	11
		C-duration (stop)	3
		C-duration (fricative)	-5
		Pulsing-duration (stop)	5*
	Fourakis & Iverson (1984) (group data)	V-duration	-4
		C-duration	-3.5
	Port & O'Dell (1985)	V-duration	15*
		C-duration	5*
		Pulsing-duration	5*
		Aspiration-duration	15*
	Charles-Luce (1985)	V-duration (__ C#V non-clause-finally)	-8(*)
		V-duration (__ C#C non-clause-finally)	-13(*)
		V-duration (__ C#V clause-finally)	-4(*)
		V-duration (__ C#C clause-finally)	-17(*)
		C-duration (__ #V non-clause-finally)	-1
		C-duration (__ #C non-clause-finally)	-1
		C-duration (__ #V clause-finally)	2
		C-duration (__ #C clause-finally)	-5
Polish	Słowiaczek & Dinnsen (1984)	V-duration	12*
		C-duration	Not given
		Pulsing-duration	13*
	Jassem & Richter (1989)	V-duration	4
		C-duration	4
		Pulsing-duration	15
	Tieszen (1997) Warsaw Polish	V-duration (__ C#V)	4
		C-duration (__ #V)	4.5
		Pulsing-duration (__ #V)	21*
		V-duration (__ C#C)	9
		C-duration (__ #C)	9
		Pulsing-duration (__ #C)	20*
	Bydgoszcz Polish	V-duration (__ C#V)	2
		C-duration (__ #V)	.6
		Pulsing-duration (__ #V)	6*
		V-duration (__ C#C)	2
		C-duration (__ #C)	8.5*
		Pulsing-duration (__ #C)	10*
	Kraków Polish	V-duration (__ C#V)	-3
		C-duration (__ #V)	6.5
		Pulsing-duration (__ #V)	5.5*
		V-duration (__ C#C)	3
		C-duration (__ #C)	-2
Pulsing-duration (__ #C)		1.5	
Dutch	Jongman (2004)	V-duration	4*
		C-duration	2
		Burst-duration	6*
	Warner et al (2004)	VV-duration	3*
		V-duration	4*
		C-duration (VV __)	1
		C-duration (V __)	2
		Pulsing-duration (VV __)	-1
		Pulsing-duration (V __)	1
		Burst-duration (VV __)	9*
Burst-duration (V __)	3		

Note: *: statistically significant; (*): statistically significant as part of an interaction between variables; +: difference in the expected direction; -: difference in the opposite direction

Table 2-1: Durational differences between 'voiced' and 'devoiced' obstruents as reported in a sample of neutralisation studies in the literature

Language	Study	Parameter	Difference (ms)		
Friulian	Baroni & Vanelli (2000)	C-duration (...a_#)	45*		
		C-duration (...o_#)	15		
		C-duration (...i_#)	70*		
		C-duration (...e_#)	68*		
		C-duration (...u_#)	26*		
Catalan	Charles-Luce & Dinnsen (1987)	V-duration (__C#V)	0		
		C-duration (__#V)	2		
		Pulsing-duration (__#V)	1*		
		V-duration (__C#C)	-1		
		C-duration (__#C)	-1		
		Pulsing-duration (__#C)	1*		
		Dinnsen & Charles-Luce (1984) (by speaker)	S1	V-duration (__C#V)	-3(*)
				V-duration (__C#C)	5(*)
	C-duration (__#V)			-7	
	C-duration (__#C)			-1	
	Pulsing-duration (__#V)			0	
	Pulsing-duration (__#C)			0	
	S2		V-duration (__C#V)	1	
			V-duration (__C#C)	3	
			C-duration (__#V)	-7(*)	
			C-duration (__#C)	-21(*)	
			Pulsing-duration (__#V)	0	
			Pulsing-duration (__#C)	3	
	S3		V-duration (__C#V)	-2	
			V-duration (__C#C)	-4	
			C-duration (__#V)	7	
		C-duration (__#C)	5		
		Pulsing-duration (__#V)	0		
		Pulsing-duration (__#C)	1		
	S4	V-duration (__C#V)	3		
		V-duration (__C#C)	6		
		C-duration (__#V)	-1		
C-duration (__#C)		-6			
Pulsing-duration (__#V)		2			
Pulsing-duration (__#C)		-2			
S5	V-duration (__C#V)	-3			
	V-duration (__C#C)	4			
	C-duration (__#V)	9			
	C-duration (__#C)	-12			
	Pulsing-duration (__#V)	3			
	Pulsing-duration (__#C)	-3			
Afrikaans	van Rooy et al (2003)	Existing words	V-duration	17(*)	
			C-duration	9*	
			Burst-duration	2	
			Aspiration-duration	14*	
		Nonsense words	V-duration	12(*)	
			C-duration	11*	
			Burst-duration	1	
			Aspiration-duration	12*	

Note: *: statistically significant; (*): statistically significant as part of an interaction between variables; +: difference in the expected direction; - difference in the opposite direction

Table 2-2: Durational differences between 'voiced' and 'devoiced' obstruents as reported in a sample of neutralisation studies in the literature [continued]

To have a feel for quantitative and qualitative variability in the phonetics of neutralisation, let us first consider the magnitude of the durational differences that are found to be statistically significant. In terms of central tendency, we find the following values (see Table 2-3 for summary statistics): the mean durational

difference that is statistically significant is 10.5ms with a standard deviation of 18ms; the mode, which is the most frequent value, is 5ms; and the median is 8.5ms. In terms of range absolute values, we can see from Table 2-1 and Table 2-2 that the smallest durational difference that has reached statistical significance is |1|ms (Charles-Luce & Dinnsen 1987), while the largest difference is |70|ms (Baroni & Vanelli 2000). In between these are durational differences varying in magnitude and in whether or not they are statistically significant. For example, a pulsing duration (i.e., the extent of the periodic signal within the stop closure) of 15ms in Polish (Jassem & Richter 1989) is not statistically significant—a finding that has been taken to support the claim that neutralisation in Polish is phonetically complete. Conversely, a pulsing duration of 13ms (Slowiaczek & Dinnsen 1984) and 6ms (Tieszzen 1997), again both in Polish, are statistically significant. These findings support the claim that neutralisation in Polish is phonetically incomplete. This situation is not peculiar to studies on Polish. We find it repeated in studies on German, Dutch, and Catalan final devoicing, for instance.

Central tendency	Mean	10.5ms (SD=18ms)
	Mode	5ms
	Median	8.5ms
Range absolute values	Maximum	70 ms
	Minimum	1 ms

Table 2-3: Summary statistics for the durational differences that reached statistical significance in studies on final devoicing

As to the directionality of phonetic differences, the mean differences in Table 2-1 and Table 2-2 clearly demonstrate how the phonetics of neutralisation exhibits all the directional possibilities that an effect may assume: positive differences, negative differences, and no differences. For example, in Port and O'Dell (1985), 'voiceless' stops are on average 5ms longer than their 'voiced' counterparts. This difference is statistically significant. In contrast, in Dinnsen and Charles-Luce (1984) 'voiceless' obstruents are in one experimental condition statistically significantly shorter by 7ms than the corresponding 'devoiced' obstruents. On the other hand, in Charles-Luce and Dinnsen (1987), we find a zero durational difference between vowels preceding 'voiced' obstruents and vowels preceding

'voiceless' obstruents. Of course, given the accepted standards within the discipline, differences that are not near statistical significance are often equated, at least as far as their inferential value is concerned, with zero differences (see chapter five for more on this).

Importantly for the current discussion, though, the interpretation of the no-difference effect is straightforward: neutralisation is phonetically complete. There is some complication, however, surrounding the interpretation of the other two effects: incomplete neutralisation entails that a phonetic difference goes in a specific direction—that which is “of the same quality as when the distinction is fully maintained” (Jongman 2004: 9). That being the case, differences that go in the opposite direction to what is expected are hard to interpret.

As is common practice in our field, a pattern of no-difference and a pattern of differences-in-the-expected-direction are treated as qualitatively distinct phonetic effects. That is, they are labelled as two different effects. The emergence of a pattern of differences-in-the-opposite-direction seems to create a dilemma for empirical researchers. Should the pattern be deemed qualitatively different from the other two 'established' patterns? What should it be called? Do we need to give priority to the *existence* of a difference over the *direction* of that difference? Or should it be the other way round? I take up these questions and the directionality issue in general in chapter five.

Also relevant is the discrepancy between the end-products of statistical tests run on group data and those run separately on data from each individual speaker. This discrepancy underscores the variability that is inherent in the phonetics of neutralisation. For example, in Gouskova and Hall (2009), statistical analyses of pooled data from eight speakers support the conclusion that vowel/zero neutralisation in Lebanese Arabic is phonetically incomplete. In contrast, analyses of individuals' data show that only three of the eight subjects actually produced a statistically significant difference between lexical and epenthetic vowels. Conversely, group data in Dinnsen and Charles-Luce (1984) show no statistically significant differences between underlyingly 'voiced' and 'voiceless' obstruents in Catalan along the durational parameters investigated. In contrast, only two of the five speakers in the study produced statistically significant differences. See chapter five for a possible explanation.

The discussion above suggests that qualitative variability can be manifested in terms of whether or not a phonetic difference exists and what direction it takes. It seems to me that the very conception of phonetically complete and incomplete effects as antitheses has contributed a great deal to the qualitative variability that we now see in the literature. In the next section, I discuss how this variability has been approached in the literature.

2.3 Approaches to Variability in the Phonetics of Neutralisation

2.3.1 Overview

Although the literature abounds with comments acknowledging the variable nature of the phonetics of neutralisation (e.g., Warner et al 2004; Gouskova & Hall 2009), they mainly refer to quantitative variability. As to variability in a qualitative sense, there seems to be a common attitude of intolerance lurking beneath verdicts such as ‘is phonetically incomplete’, ‘failed to find incomplete neutralisation’, or ‘to detect incomplete neutralisation or reliably rule it out’.

Generally, in assessing incomplete neutralisation, researchers often weigh the empirical evidence in its favour against the possible theoretical benefits of formally accommodating or dismissing it as a phonetic curiosity. Here we get hard-line as well flexible views. For example, Dinnsen (1985) asserts that neutralisation may only be phonetically incomplete—a conclusion that Manaster Ramer (1996a, 1996b) categorically rejects. Other linguists such as Barnes (2006), Jansen (2004), and Brockhaus (1995) are more sympathetic toward incomplete neutralisation. They think that the available empirical evidence actually tips the balance in its favour.

Perhaps, what all of these researchers and many others agree on is that there is a need for further investigations, which they hope might yield the long-awaited conclusive evidence on the reality of incomplete neutralisation. However, I would like to beg to differ. It seems to me that what we really need to do is to reinterpret the empirical data we already have, and of course whatever future studies yield, according to a different perspective that places emphasis on variability. As a first

step, we need to consider the various ways to deal with the issue of qualitative variability in the phonetics of neutralisation.

As we will see below, there are three major approaches to qualitative variability in the literature. These approaches involve at least five different takes on the phonetics of neutralisation listed in (6) below.

(6)

- (a) Neutralisation is only phonetically complete
- (b) Neutralisation is only phonetically incomplete
- (c) Neutralisation is neither complete nor incomplete
- (d) Neutralisation is either complete or incomplete
- (e) Neutralisation is both complete and incomplete

As mentioned above, underlying the five contentions in (6) are three different approaches to the notion of qualitative variability. They differ with respect to (1) whether they accept the complete-incomplete distinction itself and (2) whether they accept that the phonetics of neutralisation can vary in a qualitative sense. These approaches are summarised in Table 2-4.

	Distinction	Qualitative Variability	Neutralisation	Example model
Approach I	No	No	a. Only complete b. Only incomplete c. Neither complete nor incomplete	Steriade (1999) Ernestus & Baayen (2006) Gafos (2006)
Approach II	Yes	No	d. Either complete or incomplete	Barnes (2006)
Approach III	Yes	Yes	e. Both complete and incomplete	Gerfen & Hall (2001)

Note: The Gafos and the Gerfen and Hall models can be interpreted in different ways. Therefore, they can be taken as examples of different approaches (see text for more).

Table 2-4: Existing approaches to the complete-incomplete distinction and qualitative variability as relevant to the phonetics of neutralisation

In this chapter, I am more concerned with describing and illustrating the general approaches to qualitative variability—approaches that each can be seen as underlying different models of the phonetics of neutralisation in the literature. I do not attempt to give a complete review of all these models. In discussing each approach, I will present only one representative model.

As is clear from the table, approach I denies the complete-incomplete distinction altogether. If there is no distinction, there is by necessity no qualitative variability. Approach I is the general approach underlying the contentions that (a) neutralisation can only be phonetically complete, that (b) neutralisation can only be phonetically incomplete, and that (c) neutralisation should neither be phonetically complete nor incomplete, as shown in the table. Note that both (a) and (b) dismiss qualitative variability by denying the relevance of variability as well as the relevance of the distinction. In contrast, (c) dismisses qualitative variability by denying the relevance of the distinction; it still acknowledges that the phonetics of neutralisation is variable—maybe too variable to sustain a qualitative distinction.

Approach II recognises the complete-incomplete distinction but denies the relevance of variability. According to this approach, neutralisation can be phonetically either complete or incomplete. This is contention (d) in the table. In (d), we find mutual exclusivity: some neutralisations are phonetically complete; others are phonetically incomplete. Moreover, for some speakers, the effect is complete; for others, it is not. Researchers adopting this approach only need to find out which is which and for whom.

Approach III recognises both the distinction and the qualitative variability in the phonetics of neutralisation. According to this approach, neutralisation can be both phonetically complete and incomplete for the same process and for the same speakers. This is the view expressed by (e). An obvious difference between contentions (d) and (e) is that while both acknowledge the genuineness and relevance of both phonetically complete and incomplete neutralisation, (e) is less restrictive. For example, an effect that is predicted by (d) to be phonetically complete, but which is found to be otherwise, will be problematic for (d) but not for (e).

In fact, all the approaches that deny qualitative variability will have to appeal to some tool in order to explain away any variability or unexpected finding that their experimentation may reveal. So far, a strategy commonly followed is to question the genuineness of the ‘disputed’ results. For example, some researchers warn against ‘unwanted’ paralinguistic influences; others complain about the unreasonably low statistical power of many experimental studies on

neutralisation. I discuss these concerns in chapter five and show that arguments debating the genuineness of the reported findings are self-defeating, irrespective of whether they appeal to paralinguistic or statistical considerations.

Looking at the various proposals to model the phonetics of neutralisation in the literature in terms of their attitude to qualitative variability, we may be surprised to find that, despite apparent and real differences, these models are unified in their disregard of variability. As will be clear below, according to most of them, the phonetics of neutralisation is decided outside of phonetics. For example, in models assuming a feed-forward modular view of the phonetics-phonology relation, the phonological representation of the neutralised contrast determines whether we expect (and thus accept, for some linguists) phonetically complete or incomplete neutralisation. The state of the grammar will decide which pattern to expect in models as diverse as those assuming a strictly episodic exemplar-based view of the lexicon and those endorsing abstract and discrete algorithms like candidate chains optimisation.

Barnes' (2006) model of phonologisation as the phonetics-phonology interface and Gafos' (2006) non-linear dynamics stand out in this regard as the odd ones out. According to Barnes, the phonetics of neutralisation is in fact stranded between the relevant interfacing components of the grammar: processes resulting in phonetically complete neutralisation belong in the phonology, while those with neutralisation effects remaining phonetically incomplete are still in the phonetics.⁸ According to the Gafos model, phonetically incomplete neutralisation results from the interaction between grammar-dynamics and environmental (i.e., intentions) dynamics. A more fundamental difference between the two models is that Barnes' unmistakably represents approach (6)-(d), where neutralisation is phonetically either complete or incomplete. In contrast, there is some ambiguity about Gafos' model regarding how it approaches qualitative variability. Thus, the Gafos model can be taken to be a representative of any of the approaches above, even the approaches that acknowledge variability—(6)-(c) and (6)-(e). It seems to me that any interpretation the dynamics model gets depends, to a large extent, on the relevant reviewer's views of the grammar. Perhaps it is the dynamicity of the

⁸ Note that one can still argue that phonology determines the phonetics of complete neutralisation directly, whereas it determines the phonetics of incomplete neutralisation only indirectly.

model that tempts reviewers to discretise it in a way more congruent with their views of the grammar. I will discuss this model last, after I have presented all the other unambiguous approaches. But I will start with the other odd one out: Barnes' phonologisation model.

2.3.2 Either Complete or Incomplete

Barnes' phonologisation model approaches the phonetics of neutralisation with an either-or logic.⁹ Here, 'phonological neutralisation', as he calls it, is phonetically complete. In contrast, a neutralisation effect that is found to be phonetically incomplete is taken as evidence that the relevant process is only phonetic. To a hard-to-please observer, the logic behind the proposal might still be questionable: a neutralisation is phonetically complete if it is phonological, whereas a neutralisation is phonological if it is phonetically complete. Conversely, a neutralisation is phonetically incomplete if it is not phonological; and a neutralisation is not phonological if it is phonetically incomplete. Perhaps using this logic to settle research questions outside of the phonetics-phonology interface might be seen as an instance of the fallacy of begging the question.¹⁰ Another, more serious fallacy known as affirming the consequent obtains in this way:

(7) P \rightarrow Q If neutralisation α is phonetic, it is phonetically incomplete.
 Q Neutralisation α is phonetically incomplete.

\Rightarrow P Therefore, neutralisation α is phonetic.

However, in the context of the special relation between phonetics and phonology, where one can claim to see phonetics in phonology and/or phonology in phonetics, there is an element of truth and credibility in the logic Barnes uses. However, to sustain that credibility, one condition must be met at all costs: there should exist no conflict between what the phonology predicts and what the phonetics reveals, or within either of them. In situations where a conflict arises, it must be resolved in a principled way. However, given what we already know about neutralisation, we

⁹ The critique in this section only applies to the discussion of incomplete neutralisation that appears in the final chapter of Barnes (2006) and in Barnes (submitted). The thesis of phonologisation as the interface between phonetics and phonology, which takes up the rest of Barnes' (2006) book, remains one of the best approaches to the phonetics-phonology interface. For a review of this book, see Nevins (2007).

¹⁰ Barnes (submitted) acknowledges this fallacy in his account.

have no doubt that some conflict has arisen, but we have every right to doubt that it has been resolved in a principled way.

To illustrate, let us continue with Barnes' proposal, which, he admits, still has "several big 'ifs'" in it (submitted: 35). Let us rehearse once more his views: neutralisation is predicted to be phonetically complete in the wake of a process that has been phonologised. According to Barnes (2006, submitted), a phonologised process involves 'category relabelling'. In other words, the relevant contrast no longer exists in the output of the phonology. In contrast, incomplete neutralisation is an instance of phonologisation in progress, with the relevant contrast still surviving in the output of the phonology. Its neutralisation is only phonetic and thus cannot completely wipe it out. Barnes (submitted), based on experimental findings on hyperarticulated speech involving the Russian /a/-/o/ neutralisation in pretonic vowel reduction, suggests, again tentatively, that hyperarticulation might provide another diagnostic of the 'phonologicality' of a neutralisation effect. On the assumption that hyperarticulation-induced contrast enhancement can only affect the output of the phonology, goes his argument, a contrast that is neutralised in the phonology will not re-emerge, no matter how much hyperarticulation is involved in its production. Such a contrast is simply not present in the output of the phonology. Conversely, a contrast that is only incompletely neutralised can be said to be present in the output of the phonology, and thus will re-emerge rather more distinctly under conditions of hyperarticulation.

If these diagnostics are taken seriously, then we can hope to resolve the conflict issue mentioned above. At least we think we have an answer to the following question. If there is a conflict between what we think we know about the phonology of the relevant language and what we find in its phonetics, which do we trust? The answer that could be inferred from Barnes' proposal is that we should trust the phonetics. But what if the phonetics 'fails' us by giving us mixed answers? Should we accept some and reject others? What should we take as our criterion? Should what we know about the phonology of the relevant language re-qualify as a valid criterion? How should the within-phonetics conflict be resolved in a principled way?

It might be fair to say that Barnes (2006) does not suggest that we should rely exclusively on the phonetics. This will become evident in a moment. But let us first agree that in connection with resolving any precipitant conflict, Barnes' (2006, submitted) phonologisation account will have to answer this question: will investigating the phonetics of a neutralisation effect tell us if the relevant contrast is present or not in the output of the phonology? Barnes takes phonetically incomplete neutralisation "in fact, simply to be [...] evidence that [for instance] for the relevant dialects of the relevant languages, word-final devoicing is a gradient process operating in the phonetics"¹¹ (p. 227). The relevant languages that Barnes alludes to are those for which final voicing has been mostly reported to be incompletely neutralised, but for which, let us not forget, there exists a long impressionistic tradition supporting the 'phonologicality' of the process of final devoicing. These include German, Dutch, Catalan, and Polish. In other words, for Barnes, the phonetics account overrides the 'standard' phonological description of final devoicing in those languages. More generally, this means that experimental 'findings' have primacy over phonological 'facts' in his model.

COPYRIGHT MATERIAL

¹¹ This is similar to Port and O'Dell's (1985) suggestion that final devoicing in German should be 'relegated' to the phonetics for no reason other than that the phonetic investigation they ran reveals that the phonetics of final devoicing displays symptoms of incomplete neutralisation.

COPYRIGHT MATERIAL

But this is not what Barnes wishes to impart to us; nor is it all that we can find in his take on the ‘odd’ result. Barnes finds it reassuring that the difference which he believes to be ‘accidental’ goes in the opposite direction to what is expected in the case of the fully realised contrast: the schwa in place of /o/ is on average longer than the schwa in place of /a/. Perhaps the most noteworthy point he makes against ‘rushing’ to accept the effect as an instance of incomplete neutralisation, however, is when he questions the validity of the statistical significance of the durational difference. This is noteworthy not because the concern Barnes raises is necessarily valid, but because it highlights a general problem that scientific conclusions based on significance testing usually encounter (see chapter five). Barnes’ scepticism is founded on the low ‘significance level’ ($p = .049$), which he thinks “would most likely vanish in a larger study” (p. 56). Of course, this remains an empirical question. However, it is worth noting that the data Barnes gives in table (2) (reproduced below as Table 2-5) seem to have little variability both in terms of the magnitude and directionality of the difference. Accordingly, there is a realistic possibility that the difference will remain statistically significant in a larger-n study, especially that the F-associated p-value will be assessed at greater degrees of freedom.

COPYRIGHT MATERIAL

Table 2-5: Mean duration and (SD) values in ms of second pre-tonic and first pre-tonic vowels in Russian (figures based on Barnes 2006: 55)

More importantly, Barnes' protest against the genuineness of the effect above, though seemingly parenthetical, highlights once again the intolerance of qualitative variability. Put simply, Barnes' approach combines the only-complete and only-incomplete approaches into an either-complete-or-incomplete approach. This intolerance toward variability that we see in his approach is in fact inherited from the other approaches, which I present next.

2.3.3 Only Complete

For descriptive convenience, let us stick with the long-serving, descriptively useful technique of distinguishing two levels of phonological representation: (i) an underlying representation (the input to the phonology) and (ii) a surface representation (the output of the phonology). In the derivational feed-forward model, an underlying contrast is neutralised on the surface with its terms represented identically. Since the relation between the phonetic and phonological components is modular in this model, the phonetics will have no access to the underlying contrast or to what takes place in the phonology. Phonetics can only interpret what is handed over to it, which is the surface non-contrast. That is, theoretically, this model predicts that neutralisation can only be phonetically complete. For example, recall from chapter one that Kiparsky's (1973) definition of neutralisation implicates a form of realisational identity on the surface. Also, Trubezkoy's (1969) taxonomy of neutralisation includes a class where 'members of opposition' are represented identically.

Much more recently, but not necessarily assuming the feed-forward modular view of the phonetics-phonology relation, Steriade's (1999) Licensing-by-cue¹² approach to positional neutralisation predicts that neutralisation is only phonetically complete. Strictly speaking, phonetically incomplete neutralisation is incompatible with two basic tenets of Steriade's thesis. Firstly, a contrast that is incompletely neutralised is still present, if very subtly. At the same time, incomplete neutralisation is not the same as non-neutralisation. The effect is decidedly one of neutralisation, but one that is only phonetically incomplete. This does not fit well with the way licensing and neutralisation are construed in Steriade's approach. There, a contrast is either licensed, in which case we don't

¹² The discussion here only applies to Steriade's Licensing-by-cue thesis. See below for a discussion of a model that makes similar predictions as Steriade's proposal of paradigm uniformity (2000).

speak of its being neutralised; or it is neutralised, in which case we do not speak of its being licensed. In Steriade's model, contrast licensing and contrast neutralisation seem to be treated as though they are mutually exclusive: when a contrast is licensed in position X, there is no neutralisation in that position; conversely, when a contrast is not licensed in position Y, there is no contrast, but only neutralisation in that position. This follows from adopting a rather broad conception of neutralisation which subsumes both neutralisation as 'the obliteration of contrasts that exist at the lexical level' and neutralisation as 'the static lack of contrast at the lexical level'. Distinguishing between these senses of neutralisation does not seem to serve any important purpose in Steriade's study, which obviously aims to account for distributional asymmetries in the attested contrast patterns that are documented in typological studies. Steriade's paper is not a study dedicated to the phonetics of neutralisation; otherwise, such a distinction might have been warranted since these two definitions of neutralisation make different predictions regarding its phonetics. For example, incomplete neutralisation would make little sense under the view that defines neutralisation as the static lack of contrast at the lexical level.

Secondly, Steriade's account is implemented on a rigid either-licensed-or-not-licensed basis. The decision involves evaluating, among other things, the perceptibility of a certain contrast in a certain position. Steriade observes that distributional asymmetries in the typologically attested contrast patterns are directly related to the asymmetrical distribution of acoustic events that cue contrasts. All other things being equal, a contrast with diminished perceptibility is not licensed. Perceptibility here can be determined on the basis of the quantity and quality of the available cues in the relevant position. Having built her proposal on the available linguistic data, most of which are collected impressionistically, Steriade implicitly endorses the view that a licensed contrast should be robust enough to be detected by trained transcribers. This is definitely not the case with incomplete neutralisation, where the phonetic completeness of the effect has been taken for granted before instrumental analyses revealed otherwise. Ignoring, if it does, the details of the phonetics of neutralisation, Licensing-by-cue will have to treat as equally irrelevant systematic articulatory manoeuvres with imperceptible acoustic consequences and systematic acoustic perturbations with little

perceptibility—the very variations that many other linguists call incomplete neutralisation.

The alternative is not without problems either. Accepting incomplete neutralisation as evidence that the relevant contrast is licensed will create a few paradoxes for the Licensing-by-cue hypothesis. For example, if a contrast is licensed, how can we still speak of its neutralisation? What meaning will incomplete neutralisation have? Will it mean incomplete or partial licensing, for instance? Do we need to impose on the grammar (maybe the universal grammar) a distinction applying only to weak positions between those that license contrasts with naked-ear perceptibility (e.g., final /d/-/t/ in Arabic) and those that license contrasts whose perceptibility may be just above chance in speech perception tests (e.g., final /d/-/t/ in Dutch)? What cues do we need to take into consideration when we account for contrast neutralisation? Is an incompletely neutralised contrast licensed if it is imperceptible to native speakers but is perceptible to nonnative speakers? See also Barnes (2006: 230) for other arguments against Steriade's Licensing-by-cue thesis.

Returning to the only-complete approach, let us not forget that, on a practical front, there appears to be some support for phonetically complete neutralisation. The main support comes from decades of phonological tradition that is mainly based on impressionistic descriptions of speech sounds. These descriptions seem to have gained credibility by virtue of being collected and recollected by field workers with some phonetic training and whose intuitions agree most of the time. The practice of building phonological analyses and theories on data collected impressionistically rarely fail phonologists. More importantly, it has expedited the accumulation of linguistic data and knowledge which ultimately are in need of cognitive accounting. By necessity, such accounting admits of an element of abstraction and discreteness in the description of the sound patterns that phonology is mostly concerned with. This discreteness is already at the heart of the impressionistic tradition.

Moreover, the evidence for phonetically incomplete neutralisation has never been conclusive and not at all impressive either in magnitude or consistency. Broadly speaking, the phonetic difference is not always statistically significant, not always in the expected direction, and its alleged perceptibility is only above chance—a

finding that might not withstand close scrutiny. I will elaborate on these points in chapters four and five. However, I would like to suggest here that the fact that the phonetic difference is not always otherwise is also important. If its being not always statistically significant is an argument for the only-complete approach, its being not always otherwise is an argument against this very approach. Understandably, the shaky evidence for incomplete neutralisation has apparently not been enough to shake the beliefs held by the proponents of the only-complete approach about the phonetics of neutralisation. Yet, these researchers are definitely not adamantly opposed to subjecting the phonetics of neutralisation to further experimentation with improved methodology, so as not to repeat the 'old mistakes' (see e.g., Manaster Ramer 1996a, 1996b).

2.3.4 Only Incomplete

When the phonological community was ready to embrace incomplete neutralisation, it actually overdid its welcome. Phonologists have come up with a number of proposals which differ in technicalities, but all serve a common purpose—to formally accommodate and predict incomplete neutralisation as a genuine grammatical effect. However, most of their accounts, while successfully predicting phonetically incomplete neutralisation, predict that neutralisation cannot be otherwise. This is the only-incomplete approach. Proponents of this approach typically accuse studies reporting complete neutralisation of having low statistical power. They explain that the effect is certainly there, but that the researchers have not done enough to find it.

It seems to me that there may be two distantly related causes behind phonologists' belated enthusiasm for incomplete neutralisation—one relevant to competing approaches to the phonetics-phonology relation, the other to phonologists' general approach to their field of study. Firstly, the recent revelation that systematic fine phonetic detail has a role to play in speech perception (e.g., Smith 2004; Stager & Werker 1997), word-recognition (e.g., Hawkins 2003; Hawkins & Nguyen 2003), and infant language processing (e.g., Johnson & Jusczyk 2001) warrants a re-evaluation of the way the grammar has been conceived of in the past. Researchers have called for fine-grained phonetic effects to be integrated into a new approach to the lexicon. It has been claimed that lexical properties (syntactic, etymological, statistical, etc) are instantly transferable not only into phonology, an idea which

has been around for a while, but also into phonetics. In addition to recognising a phonology of lexical strata (e.g., Giegerich 1999; Itô & Mester 1999), loanword phonology and native phonology (e.g., Calabrese & Wetzels 2009; Kang, to appear) verb phonology and noun phonology (e.g., Smith 2001; Lee 2001), researchers have also argued that language-specific, speaker-specific, style-specific, word-specific, and phone-specific phonetics (see references below) must all be integrated into the high-level components of the grammar.

Naturally, this move has serious consequences for the phonetics-phonology relation. For example, it may involve serious blurring of the long-standing boundary between phonetics and phonology by making the phonological grammar more phonetics-conscious. This can be achieved by constructing a hybrid component of the grammar whose primitives can have both discrete, or more traditionally 'phonological', values *and* continuous, or 'phonetic', values. The celebrated lack of acoustic invariance in speech production might force the concession that there still exists a separate non-phonological phonetics. The continuous values the hybrid component contains will have to remain idealisations that non-phonological phonetics may or may not realise (cf. Nguyen et al 2009). That being the case, we can see that at least the non-phonological phonetics is still dependent for its input on the hybrid phonology. Note that a strong version of the commitment to fine phonetic detail can totally abolish the distinction between phonetics and phonology by banishing this low-level, real-time, and non-idealised 'phonetics' from grammar proper. On this extreme view, then, non-idealised phonetics will be treated as a peripheral neuro-physiological system interfaced not with the hybrid component of phonetics-phonology but with the grammar as a whole.

Another radical move is to keep the distinction between phonetics and phonology but reverse the direction of the dependency relation. This is best exemplified by the exemplar-based view of the phonetics-phonology interface (e.g., Goldinger 1997, 1996; Pierrehumbert 2001; Lachs et al 2003; Johnson 1997, 2007). Production, according to this model, consists in accessing a sub-space of a cloud of tokens of speech memorised with their full phonetic profiles. The selection of a specific sub-region to access is actually biased by various effects including co-activation of morphologically related forms, recency, frequency, etc. The selection

of a target for production can follow a statistical averaging of many exemplars within the activated space or some randomisation process. When a new phonetic rendition of a unit of speech that clearly belongs to a certain cloud is encountered, it is added to the exemplars populating that cloud. Almost concurrent with this addition is an updating process whose outcome is not predicted to be huge after the addition of one exemplar, unless the new addition is an extreme rendition. Normally, the difference due to an addition will almost always be averaged out, and so there will be neither a change in the phonetic distribution nor category re-labelling. In this way, speakers belonging to the same linguistic community will converge on very similar values. This allows for generalisations to be made and supported.

However, the phonetic distribution of a cloud of exemplars can change in shape and size. For example, Yu (2007) illustrates how the phonetic distribution of a single category label may overlap another cloud defining another category label to such an extent that there are not enough phonetic differences between the two to support a category distinction. Conversely, a phonetic distribution of exemplars representing a single category can develop bi-modality where averaging occurs in two different sub-regions at a consistent rate. The result would be that two different labels may be defined over these sub-regions. The first scenario, according to Yu, underlies complete mergers; the second represents what is known as split. Incomplete neutralisation or near mergers are similar to the first scenario, except that the overlap is not total. Yu's account of mergers and near-mergers shares with Barnes' (2006) phonologisation model the underlying either-or logic. For example, once phonologisation or merger is complete, a phonetics of incomplete neutralisation or near-merger is highly suspicious within the synchronies of the relevant languages. However, in Yu's account a quick reversal of a complete neutralisation is not unlikely. Here, phonology seems to be more parasitic on the phonetic addition and removal of exemplars, which in turn trigger statistical updating. As far as the theory goes, there is nothing that prevents the phonetic distribution associated with a merged label from eventually developing bimodality. Such bimodal distribution may in time split and become associated with two different category labels, reversing the merger or giving rise to some altogether new categories. But again this takes time, and so, within a short span of

the synchronic grammar of a speaker, the phonetics of neutralisation will have to display either a complete or an incomplete effect.

Although Yu's exemplar-based account is a model with an implicit either-or approach to qualitative variability, its direct application to incomplete neutralisation (e.g., Ernestus & Baayen 2006) can only derive an only-incomplete effect. This directly follows from the morpho-phonological structure of the speech material that is subject to neutralising processes. In almost all of those neutralisation cases whose phonetics has been scrutinised, the test items are words with relatives fully displaying the relevant contrast. The main idea of Ernestus and Baayen's proposal is that these morphological relatives can actually bias the production of the items where a contrast is supposed to have neutralised, thus resulting in incomplete neutralisation. This is due to the co-activation of these words. For example, the co-activation of a word ending in [t] in Dutch and its relative whose corresponding sound is [d] will bias the production of the [t] towards being more [d]-like. This is an elegant way of capturing Steriade's (2000) notion of paradigm uniformity. Producing a [t] in a more [d]-like fashion as a result of what Barnes (2006: 233) aptly calls "morphophonetic gravitation" would be the normal phonetics of the voicing neutralisation in Dutch. As long as there are morphological relatives displaying a certain sound contrast, neutralisation of that contrast can only be phonetically incomplete. Since the lexical biasing influence responsible for incomplete neutralisation is an automatic reflex of the grammar,¹³ declaring its absence, as would be understood by reporting a complete neutralisation, would just go against the essence of the exemplar-based view of the lexicon and the phonetics-phonology relation.

However, it seems to me that the model makes no provision for why morpho-phonetic gravitation should go in a specific direction, or more generally, for why we should not expect the biasing effect to be bi-directional between the base and derivative. Bi-directionality can have very serious consequences for the phonetics of neutralisation, at least conceptually speaking. For example, according to Ernestus and Baayen (2006), the production of the uninflected Dutch word [vɛrʋeɪt] 'widen' is biased towards having a 'slightly voiced' final sound, as a result of the lexical co-activation of its inflected relative [vɛrʋeɪdɔn] 'to widen'. They

¹³ A similar idea can be found in Snoeren et al (2006).

argue for a complete correspondence between lexical representations and pronunciation. For example, “[t]he form *verwijd* needs not be stored as /*vɛrvɛid*/, but can be stored as /*vɛrvɛit*/, directly reflecting its pronunciation” (p. 46). Note that there is nothing in the model that should a priori invalidate the reverse situation. In fact, the logic of their argumentation could equally plausibly apply in the opposite direction, where the production of [*vɛrvɛidən*] is biased towards slight voicelessness by its simplex base [*vɛrvɛit*].¹⁴ At least, it seems reasonable to accord more privilege to the base than to its morphological derivatives. The phonetic and phonological literature provides more empirical and typological evidence in favour of the primacy of the base (e.g., Kenstowicz 1996; McCarthy & Prince 1995; Steriade 2000).

Now if we argue that there is nothing that should restrict the directionality of the biasing effect, we shall need to explain why, when biasing effects can go either way, our model does not predict that, at some point, opposing effects will cancel each other out. More concretely, suppose that the production of the [t] in [*vɛrvɛit*] is biased towards sounding more [d]-like by the existence of its [d]-relative [*vɛrvɛidən*]. At the same time, the production of the [d] in the latter is biased towards sounding more [t]-like, by the existence of the former. If this mode of phonetic rendition, which would eventually join the respective exemplar-clouds of the respective categories, persists for some time, what would be left of the more [t]-like [d] to make its more [d]-like [t]-relative more [d]-like? Of course, we could stipulate that biasing only occurs on-line and that its effects are short-lived and momentary. But that should undermine the episodic spirit of exemplar models. Now as far as this situation is concerned, an appeal to the notion of phonetic naturalness may salvage Ernestus and Baayen’s account, as voicing is expected in intervocalic context (see e.g., Westbury & Keating 1986). Yet, phonetic naturalness may not always be available to complement their model.

More relevant to the issue of variability is the fact that the model will have to provide answers to the following questions. Does the inter-speaker variability that we see in experimental data represent the rendition of a phonetic target within a

¹⁴ This is reminiscent of Steriade’s (2000) analysis of ‘cyclic’ effects involving correspondence between a base and its allomorphs. The driving force behind that correspondence, according to Steriade, is morpheme invariance or paradigm uniformity.

single category cloud averaged differently by the different participants in an experiment, or does it represent the rendition of targets belonging to different category clouds? Is it not plausible to imagine a situation where most speakers target a realisation of, say, [t] within the t-category cloud, while a few target the realisation of [d] within a d-cloud? In other words, how can we rule out the possibility that the subjects in an experiment actually have different category labels (cf. Yip 1996)? Similarly, does the intra-speaker variation we see in experimental data result from different pools for averaging within the bounds of the same cloud for the relevant speaker?

Practically, there is ample evidence for the relevance of low-level sub-phonemic effects to speech processing and production (see above for references). However, these effects certainly add to the naturalness and nativeness of the spoken message and therefore speed up word-recognition. Yet for communicative purposes, sub-phonemic effects might only be of secondary importance. Consider how badly an orally transmitted message is affected when stripped of all sub-phonemic information but which is presented in good listening conditions. Most probably, it might be judged as unnatural, machine-like, or foreign-accented speech. However, it would be comprehended fully. At least, this would be the performance by those who speak the relevant language natively, or near-natively. With beginning non-native learners, however, the situation is different. Here, long exposure to the target speech with full-scale variability is what leads eventually to an adequate comprehension of non-native speech (see chapter six for details). The Variability Field model sketched in this thesis hypothesises that language acquisition involves learning about the nature and bounds of phonetic variability in speech. Only after sufficient exposure to variability do we, as speakers-hearers, reach a stage where we have made many generalisations and associated with each a certain margin of freedom that each one of us can exploit while orally communicating that generalisation to others and to ourselves. If the communicated generalisation falls outside of that margin, we either reject it as being foreign, unnatural, or weird, or accept it but modify the margin we have previously associated with it. The latter is the general strategy of non-native learners. In contrast, native speakers follow different strategies depending sometimes on how distant the stimulus is from the margins. I present a detailed sketch of this model in chapter six.

As I pointed out above, another reason for phonologists' belated interest in incomplete neutralisation concerns their own approach to their discipline's expressed mission. It might not be an exaggeration to say that the enthusiasm for fine phonetic detail on the part of many empiricists and functionalists has made phonologists only too aware of a retreat in the appeal of formal phonology. These phonologists plead that phonological machinery may still be capable of handling sound variations exhibiting systematicity that is not readily exclusively of a phonetic flavour. Of these, variation whose systematicity is non-automatic, goes the argument, deserves to be treated as a higher-order effect whose rightful place is in the phonology. One such challenging phenomenon for formal phonology to model is incomplete neutralisation. van Oostendorp (2008: 1372) maintains that

Formal phonologists thus need to take these facts [incomplete neutralisation] seriously, and try to incorporate them into their model of phonology. The more conservative approach is to *not* give up the whole enterprise of formal analysis in the face of a few problematic data [italics his].

van Oostendorp (2008) offers a formal phonological account of incomplete neutralisation in German final devoicing. His proposal rests on theoretical assumptions and tools from Turbidity Theory cast within an OT framework. To van Oostendorp, incomplete neutralisation is phonetics reflecting phonologically distinct representations of the supposedly neutralised contrast. Phonetics does this by obeying purely phonological laws in not realising certain phonological material in positions of neutralisation. This unrealised material is a pronunciation relation that the feature [voice] cannot have with final obstruents. The proposal makes the following prediction: a devoiced obstruent, say final /d/, should be phonetically more similar to its corresponding voiced counterpart, i.e., /d/ elsewhere, than to the corresponding voiceless sound, i.e., /t/. This follows from the fact that a devoiced obstruent is representationally more similar to its voiced counterpart in van Oostendorp's account. This is schematised in (i) in (8) below. Contrast this with the following surface representations from the other more traditional models in (ii) and (iii) below.

(8)

COPYRIGHT
MATERIAL

The three models above make different predictions regarding the phonetics of neutralisation. For example, (ii) predicts that neutralisation may only be phonetically complete. Conversely, both (i) and (iii) predict that neutralisation may only be phonetically incomplete. Importantly, (i) and (iii) also predict that the devoiced obstruent (e.g., final /d/) is more similar to the voiced obstruent elsewhere (e.g., /d/ elsewhere) than it is to its voiceless counterpart (e.g., /t/). But this prediction is simply false. Even in situations where acoustic differences between final ‘voiceless’ and final ‘devoiced’ stops in a language like Dutch reach statistical significance and are in the ‘expected’ direction, they can never be comparable to the acoustic differences between final ‘devoiced’ stops and the corresponding ‘voiced’ stops elsewhere. These latter differences will not only be always of the expected quality but will also consistently have an impressive magnitude. These models code a symbolic near-identity between devoiced

obstruents and their voiced counterparts—a position that phonetics does not support.

2.3.5 Flexible Models

I discuss here two models: Gafos' (2006) non-linear dynamics and Gerfen and Hall's (2001) proposal of relaxing the strict modularity assumption in the phonetics-phonology relation. Let me point out, however, at this early stage of the discussion, that the flexibility of these models in dealing with the phonetics of neutralisation is only apparent. More specifically, it is their amenability to disparate interpretations that makes them look flexible. Note that this is not the case with the strictly modular feed-forward view, which is truly compatible with all the three approaches described above: approach (6)-(a), where neutralisation can only be phonetically complete; approach (6)-(b), where neutralisation can only be phonetically incomplete; and approach (6)-(d), where neutralisation can be phonetically either complete or incomplete. Within the strictly modular feed-forward scheme, what differentiates these approaches is whether or not the relevant contrast is allowed to make it to the output of the phonology, which, and nothing else besides it, is inputted to the phonetics. Approach (6)-(a) will allow no contrast to be represented in the output of the phonology; in approach (6)-(b), there will always be a contrast to be handed over to the phonetics; and approach (6)-(d) will foster some selectivity in resolving the question of which is which. But, that would not be more than an exercise of choosing between (6)-(a) and (6)-(b). Models assuming a feed-forward modular view do not lend themselves to disparate interpretations as to their approaches to variability in the phonetics of neutralisation. These models are not flexible in the sense adopted here. Of the two flexible models I discuss here, however, Gerfen and Hall's (2001) comes in a formal language that is more familiar to all phonologists. I consider this model first.

According to Gerfen and Hall (2001), incomplete neutralisation obtains when the phonetics accesses the underlying distinction between the relevant sounds.¹⁵ In other words, there are two 'sources of information' for the phonetics: the input to the phonology as well as the output of the phonology (see Figure 2-1 below). Thus,

¹⁵ This is very similar to Gouskova and Hall's (2009) account of incomplete neutralisation in Lebanese Arabic. However, the modularity is upheld in the Gouskova and Hall account, which employs Candidate Chains in OT (e.g., McCarthy 2007a). According to their model, the underlying distinction (the input) is already handed over to the phonetics as part of a winning output candidate chain.

the relation between phonetics and phonology cannot be assumed to be strictly modular. This, they observe, is a middle-ground position between two extremes: rejecting the distinction between phonetics and phonology or rejecting incomplete neutralisation.



COPYRIGHT MATERIAL

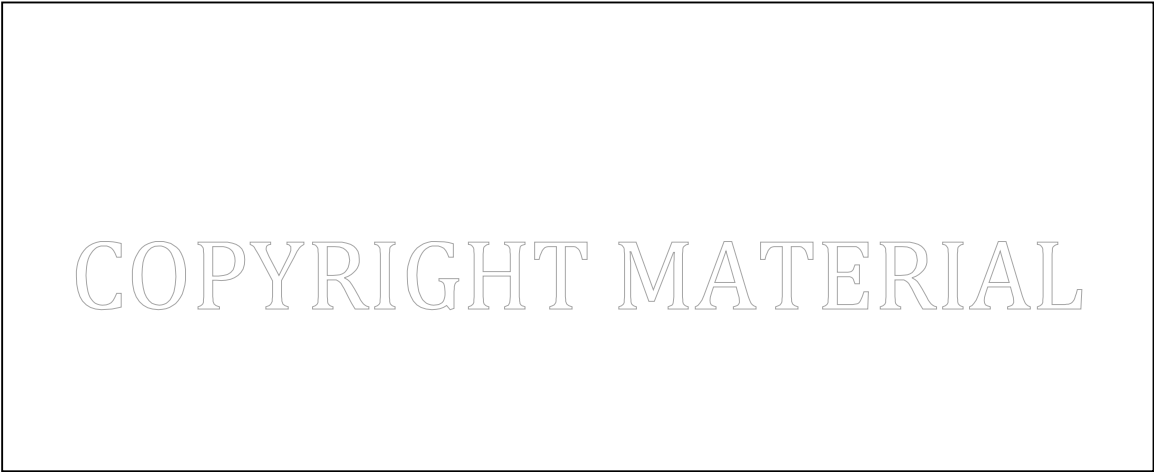
Figure 2-1: Phonetics accessing the input to the phonology (reproduced from Gerfen & Hall 2001: 31)

How can Gerfen and Hall's proposal be interpreted differently to represent more than one approach to qualitative variability? If the sensitivity of the phonetics to underlying distinctions is made mandatory by some universal mechanism, for instance, then we can only expect the phonetics of neutralisation to display signs of incompleteness. This is approach (6)-(b) where neutralisation can only be phonetically incomplete. Conversely, the either-or scenario (i.e., approach (6)-(d)) is also possible: when the phonetics is granted access to the underlying distinction, we get incomplete neutralisation; when that access is denied, however, we get complete neutralisation. But an obvious weakness of this latter interpretation stems from its potential for circular argumentation. For example, one can assume that incomplete neutralisation results from this non-modular relation between phonetics and phonology, and provide as evidence for the existence of such an open-door relation incomplete neutralisation itself. One way around this circularity, not necessarily effective, though, is to have the antecedent-consequent order inversed: incomplete neutralisation is phonetics accessing the underlying distinction, whereas complete neutralisation is what we get when the phonetics does not see the input. Note that the different phrasing of what is essentially the same idea affects how strongly we may react to the circularity of this logic. This may partly explain why the recurrent conjecture that incomplete neutralisation is

phonetics delving deeply into the phonology has never been pre-emptively questioned on logical grounds.

If, while working within the general framework of Gerfen and Hall's (2001) model, we adopt approach (6)-(d) (i.e., neutralisation is either complete or incomplete), we will have a few moot questions to resolve. These questions arise when we get results that go against our expectations. We ask then why the phonetics does not access the phonology when it should, and why it does when it should not. Adopting approach (6)-(e), where neutralisation can be both complete and incomplete, instead, we will have an altogether different set of questions to answer. Here, both effects of complete and incomplete neutralisation are tolerated. A possible set of questions that we might have to answer is why the phonetics does not access the phonology when it does not, and why it does when it does.

The next flexible model I discuss in this section is Gafos (2006). This model appeals to concepts of the mathematics of non-linear dynamics to explain incomplete neutralisation. The essence of the account is to formalise the online derivability of incomplete neutralisation from the interaction of grammar and intentions. As far as incomplete neutralisation is concerned and simplifying matters a bit, the intention to convey a contrast conflicts with a grammar requirement to suppress that contrast. However, in this model, these opposing pressures are not modelled as static phonological constraints as in main-stream OT, for example. The dynamicity accorded to these constraints is a natural consequence of their being conceived of as competing attractors in "a numerically-defined multidimensional state space" (Nycz 2005: 277) as illustrated in Figure 2-2.



COPYRIGHT MATERIAL

Figure 2-2: The dynamical models of two grammars; (x) is a continuous articulatory variable (Gafos 2003a: 8)

Gafos has modelled final devoicing in German. The phonetic data he is concerned with come from a series of experiments reported in Port and Crawford (1989). The experimental design of their study includes various speech tasks (experimental conditions) presumably placing different demands on the participants to pay attention to the test materials. These conditions include reading and repeating minimal pairs embedded in semantically composed sentences providing disambiguating clues, dictating minimal pairs in frame sentences to a waiting assistant, and reading a list of minimal pairs in isolation. The degree of incomplete neutralisation was found to be different among the different conditions. Gafos (2006) uses this finding to argue for the “subtle dependency [of incomplete neutralisation] on the communicative context” (p. 54). The more contrast-promoting the experimental condition, the greater the incompleteness of neutralisation we observe. For example, the condition where participants dictated a list of sentences to a waiting research assistant induced a greater degree of incomplete neutralisation than was found in the condition where participants were merely reading out a list of randomised words. Gafos took this context-dependent increase/decrease in the degree of incomplete neutralisation to be directly related to the strength of the communicative intention on the part of the speaker to ‘convey the contrast’. More formally, when the intention dynamics throw up an attractor towards the voiceless end of the state space, and the grammar dynamics throw an attractor at the same part, the grammar-intention interaction is said to be an instance of cooperation. This is the case when a German speaker wants to convey a [t] word-finally. In contrast, when the intention of that speaker is to convey [d], there is a competition between this intention and the grammar, which requires final stops to be devoiced. It is due to this competition, according to Gafos, that we get incomplete neutralisation. The stronger the intention for [d], the more [d]-like the relevant sound will sound, and the greater the incompleteness of the neutralisation between final /t/ and /d/ we will observe (see Figure 2-3).

COPYRIGHT MATERIAL

Figure 2-3: Cooperation and competition between grammar dynamics and intention dynamics in the production of final 'voiced' and 'voiceless' obstruents in German (reproduced from Gafos 2003a: 16-17)

To Gafos, the variability in the magnitude of incomplete neutralisation is “lawful continuous variation” (p. 57). This notion can also extend to complete neutralisation. In this way, Gafos’ model could be said to represent approach (6)-(c), where the phonetics of neutralisation is just too variable for the complete-incomplete distinction to have meaning. This conclusion is confirmed by the fact that the model is capable of over-prediction: increasing the intention for [voiced] “beyond some relatively high value [...] [t]he model then predicts a bifurcation, a qualitative change in the system’s dynamics” (p. 72), where the voicing contrast fully re-emerges. How lawful is this qualitative change? Gafos does not see this as a problem but observes, instead, that a native speaker of German is able to produce a fully voiced [d] word-finally if he/she is willing enough. However, the argument also holds for predicting the realisation of final /t/ as fully voiced [d] by a German speaker who has the intention to do so. And how this is done is simply a matter of intentions overriding the grammar. The important question then is whether or not such weird pronunciations should be considered ‘grammatical’. If they are not grammatical precisely because their production involves suppressing the grammar in favour of conveying anti-grammar intentions on the part of the speaker, then incomplete neutralisation must also be seen as belonging to this group. Gafos suggests in passing that intentions are constrained by “how forms ‘should be produced’ in specific context” (p. 67). But this is part of introducing the notions of cooperation and competition between grammar dynamics and intention dynamics.

In this context, the grammar does not really constrain what communicative intentions a speaker may have. But the level of gratification of intentions is dependent on whether there are full blessings from the grammar, or if there is a bit of a conflict. In the latter case, the conflict will eventually have to be resolved in favour of one of the conflicting parties. That is, there will occur intentionally weird pronunciations like a non-native [θ] in place of a native /t/ unless we define the range of values along which the acoustic/articulatory variable (x) can vary. A less attractive alternative is to introduce a mechanism into the model that can filter what intentions can legally interact with the grammar.

Gafos' model can also be interpreted in either-or terms. The model predicts incomplete neutralisation in speech produced and recorded as part of a laboratory study, where participants are presumably more conscious of what they say and how they say it. In settings where intentions do not need to be communicated to be known, as when one speaks to himself/herself, the model predicts that neutralisation can only be complete since verbally communicating the intention to convey a contrast to oneself by producing more of it is redundant at best and counter-intuitive at worst.

On the other hand, Gafos' characterisation of incomplete neutralisation as an instance of intentions modifying the grammar can translate into para-linguistics influencing linguistics, particularly for linguists whose research vocabulary includes such items. To these researchers, any such interactions are epiphenomenal to the study of the grammar. On this view, an adequate study of the grammar should only be concerned with predicting and modelling complete neutralisation. In other words, Gafos' non-linear dynamics can also be used, appropriately or not, to reinforce an only-complete approach to qualitative variability in the phonetics of neutralisation.

2.3.6 Summary

As we have seen, different models of the phonetics of neutralisation make different predictions regarding the complete-incomplete distinction. Some predict that neutralisation is complete; others predict that it is incomplete; yet others predict that it can be both complete and incomplete. With reference to the last category, a distinction can be drawn between models that predict complete and incomplete

neutralisation for the same phonological process, and those that predict complete neutralisation for some processes and incomplete neutralisation for others. Finally, there are models assuming that the phonetics of neutralisation is so variable that a distinction between complete and incomplete neutralisation cannot be drawn reliably.

2.4 Conclusion

The chapter has reviewed the literature on the phonetics of neutralisation. The focus has been on the issue of variability. The chapter has examined various quantitative and qualitative aspects of the phonetics of neutralisation and suggested that the phonetics of neutralisation is variable. Quantitatively, there is a lot of variation in the magnitude of the phonetic differences between sounds that are subject to neutralisation as reported in the literature. Qualitatively, variability is seen in terms of the presence and directionality of statistically significant differences. Some studies found a statistically significant difference that has escaped other studies investigating the same neutralisation effect in the same language. The consequence of this is that the same neutralisation case has been classified as incomplete on the basis of conclusions drawn from studies revealing statistically significant differences, but as complete on the basis of conclusions drawn from studies reporting no statistically significant differences.

The chapter also offered a critical evaluation of the existing approaches to qualitative variability and the complete-incomplete distinction. These approaches fall into five groups with respect to their underlying conception of the phonetics of neutralisation. One group only recognises phonetically complete neutralisation as both genuine and relevant; another group only predicts phonetically incomplete neutralisation; a third group is a combination of the first two, with complete neutralisation accepted in certain cases, and incomplete neutralisation in others; a fourth group predicts that neutralisation can be both phonetically complete and incomplete for the same phonological process and the same group of speakers; and, lastly, a fifth group assumes that the phonetics of neutralisation is so variable that a distinction between complete and incomplete neutralisation cannot be drawn reliably.

3 The Phonology of Vowel/Zero Neutralisation in BHA

3.1 Introduction

In Bedouin Hijazi Arabic (BHA), the underlying contrast between the presence and the absence of a vowel is putatively neutralised. This chapter describes the phonology of this putative neutralisation, which I have previously referred to as vowel/zero neutralisation. I will continue to use this term throughout the rest of the thesis. The phonological process that is responsible for this neutralisation effect is vowel epenthesis, and it is in focus here.

The peculiarities of epenthetic vowels as contrasted with lexical vowels are well-documented in the phonological literature (see e.g., Archangeli 1988, 1984; Archangeli & Pulleyblank 1994; Broselow 1982; de Lacy 2002a; Kitto & de Lacy 1999; Lombardi 2002). For example, in some languages, epenthetic vowels behave differently from their corresponding lexical vowels with respect to *stress assignment* (see e.g., Alderete 1996; Alderete & Tesar 2002; Łubowicz 2003; Piggott 1995), *assimilation* (e.g., Herzallah 1990), *vowel raising* (McCarthy 2007b, 2003), and *vowel harmony* (Finley 2009, 2008; van Oostendorp & Revithiadou, in progress). The sensitivity of these (and other) phonological processes to whether the relevant vowel is epenthetic or lexical provides a rich source of opacity in the world's languages (see chapter one for more on this).

To account for the opacity problem, phonologists have come up with formal devices such as *rule-ordering* (SPE and generative tradition), *Sympathy Theory* (McCarthy 2003, 1999), *Candidate Chains* (McCarthy 2007a, 2006), *Local Conjunction* (Smolensky 2006; Łubowicz 2002; Itô & Mester 2003), and *Turbidity Theory* (Goldrick 2001; Goldrick & Smolensky 1999). However, even though

phoneticians and laboratory phonologists have been engaged in scrutinizing phonological phenomena involving neutralisation for decades, contrasts that are putatively neutralised through epenthesis rarely figure in their studies. When phonetic studies started to yield results that were at odds with impressionistic descriptions (see chapter two), one would have thought that neutralising vowel epenthesis would have made an obvious choice, given the opacity problem mentioned above.

Consider for instance the following scenario: a lexical vowel and an epenthetic vowel which are impressionistically described and transcribed as the same in terms of both quality and quantity and which occur in the same environment are in fact pronounced differently by native speakers. That is, the underlying vowel/zero contrast is not completely neutralised, and the problem of epenthesis-related opacity is only the phonologist's.

Despite its relevance to the incomplete neutralisation issue, vowel epenthesis has attracted little attention. In fact, only a few¹⁶ phonetic studies of this process exist (e.g., Gouskova & Hall 2009, 2007 for *Levantine Arabic*; Cristófar-Silva & Almeida 2008 for *Brazilian Portuguese*; Smorodinsky 2002 for *English*).

Gouskova and Hall's (2009, 2007) study is the most relevant to this thesis. In more detail, it is an acoustic study of vowel epenthesis in two Levantine dialects of Arabic that differ with respect to the capacity of their epenthetic vowels to bear word stress. Epenthetic vowels receive word stress more often in Palestinian Arabic than in Lebanese Arabic, where an epenthetic vowel is only stressed when it breaks up a cluster of four consonants. See (9) and (10) for illustrations.¹⁷

¹⁶ Of course, including studies of the phonetic properties of intrusive and excrescent vowels (see below for references) would significantly add to the size of our collection. However, it must be noted that these vocalic events differ in fundamental respects from epenthetic vowels. For example, whereas vowel epenthesis repairs marked structures (e.g., final clusters with a rising sonority profile, clusters violating the Obligatory Contour Principle (OCP), etc.), vowel intrusion occurs in well-formed structures. Moreover, the presence or absence of an intrusive vowel within a consonant cluster does not trigger, block, satisfy, or violate any phonological processes or constraints including template constraints (for more on intrusive vs epenthetic vowels, see Hall 2003, 2006; Buchwald 2005; on excrescent vs epenthetic vowels, see Levin 1987; Hall 2003; Campbell 1974). In Almihamdi (2006), I present empirical data supporting the claim that BHA has both vowel epenthesis and vowel intrusion, and that these are unambiguously distinct. In this thesis, I am only concerned with vowel epenthesis in BHA.

¹⁷ Data are from Farwanah (1995: 46) and Gouskova and Hall (2009: 204-206, 209). I have verified the Palestinian items with a native speaker.

(9)

COPYRIGHT MATERIAL

(10)

The data in the Gouskova and Hall study were collected from eight native speakers of each dialect reading out a list of /CVCVC/ and /CVCC/ minimal and near-minimal pairs. The test materials in the study mixed obligatory and optional epenthesis,¹⁸ with the consequence that some speakers failed to produce epenthesis in some test words.¹⁹ The results of the acoustic study Gouskova and Hall have conducted suggest that neutralisation is phonetically complete in Palestinian Arabic, but that epenthetic vowels are shorter and backer in the production of some of the speakers of Lebanese Arabic. Tests of statistical significance run on the pooled data from the eight Lebanese speakers suggest that neutralisation in Lebanese Arabic is acoustically incomplete.

With the Gouskova and Hall study, we need to keep in mind that there are occasions where epenthetic vowels in Palestinian reject stress just as there are occasions where epenthetic vowels in Lebanese accept stress. Put differently, epenthetic vowels in Palestinian are not always stressable; nor are epenthetic vowels in Lebanese always non-stressable. It is probably true to say that the underlying vowel/zero contrast in Palestinian Arabic is completely neutralised in the phonology except in those few cases where it still refers to the underlying contrast on the surface, as when epenthetic vowels, unlike lexical vowels, reject

¹⁸ In Levantine Arabic, epenthesis obligatorily breaks up final Obstruent-Sonorant clusters (e.g., /ʒisr/ → [ʒisir] 'bridge'). But in some final Obstruent-Obstruent clusters or Sonorant-Obstruent clusters, epenthesis is optional (e.g., /libs/ → [libis]~[libs] 'clothes'; /ʔird/ → [ʔirid]~[ʔird] 'monkey'). See Gouskova and Hall (2009, 2007).

¹⁹ Gouskova and Hall (2009) seem to have anticipated this. They included a backup list of rhyming words to use for statistical analysis if speakers failed to produce the correct form of the target words in the original list. However, it seems to me that a better strategy would have been to limit analysis (or data collection) to words where epenthesis is required.

stress. By the same token, it is probably true to say that the underlying vowel/zero contrast in Lebanese Arabic is incompletely neutralised in the phonology except in those few cases where epenthetic vowels receive stress, just like lexical vowels. As a result, the phonological completeness–incompleteness of the neutralisation pattern in both Levantine dialects is complicated by the existence of these ‘exceptions’ in the stressability of epenthetic vowels as opposed to lexical vowels.

The case-study I report in this thesis has at least two advantages over the Gouskova and Hall study. Firstly, vowel/zero neutralisation in BHA delivers no such ‘exceptions’ (for illustration see §3.2.2). For example, when epenthetic [a] occurs in a stress site, it invariably receives stress, just like lexical /a/. Conversely, when epenthetic [i] occurs in a stress site, it invariably rejects stress, unlike lexical /i/. The consequence for the phonological completeness or incompleteness of the vowel/zero neutralisation pattern in BHA is this. [a]-epenthesis results in phonologically complete neutralisation, whereas [i]-epenthesis leads to a neutralisation effect that is not phonologically complete.

A second advantage of this study is that it offers to explore the phonetics of these neutralisation effects using data that are generated within a single experimental paradigm, produced by the same speakers, and analysed following the same procedures. Test materials are minimal pairs of words with epenthesis breaking up word-final Obstruent-Sonorant clusters only (see §3.2.2 and chapter four for more details).

The exploration of neutralisation in this study raises interesting questions for the phonetics-phonology relation. For example, it allows us to ask if the surface phonology can insist on an underlying contrast for which the phonetics has no expression, and conversely if the phonetics can still support an underlying contrast that the surface phonology has given up on. These are some of the questions I address in this thesis.

The rest of this chapter proceeds as follows. In §3.2, I provide background information about the dialect (BHA) and vowel epenthesis in BHA. In §3.3.1, I briefly describe a tentative conceptualisation of vowel/zero neutralisation and how vowel epenthesis neutralises underlying vowel/zero contrasts. In §3.3.2 and §3.3.3, I illustrate the opacity that characterises the interactions of vowel

epenthesis with stress and vowel-reduction processes in BHA phonology. In §3.4, I conclude the discussion by briefly touching on a prediction made by the phonology of vowel/zero neutralisation in BHA for its phonetics.

3.2 Vowel Epenthesis in BHA: Background

3.2.1 *The Dialect: BHA*

BHA is a Bedouin dialect of Arabic spoken by the Harb tribe²⁰ in the Hijaz Province in the west of Saudi Arabia. The phonology of this dialect is described in Al-Mozainy's 1981 PhD thesis. Table 3-1 and Table 3-2 give the phonemic inventory of BHA (based on Al-Mozainy 1981):

	t/t:	t ^h /t ^h :		k/k:		ʔ/ʔ:
b/b:	d/d:			g/g:		
	f/f:	θ/θ:	ʃ/ʃ:	x/x:	ħ/ħ:	h/h:
		ð/ð:	ð ^h /ð ^h :	ʕ/ʕ:	ʕ/ʕ:	
		s/s:	s ^h /s ^h :			
		z/z:				
m/m:	n/n:					
	l/l:	r/r:				
			j/j:			w/w:

Table 3-1: Consonant inventory of BHA

i/i:	u/u:
a/a:	

Table 3-2: Vowel inventory of BHA

With such an impoverished vowel inventory, it seems paradoxical²¹ that it is the vowel system that has kept the dialect's name in the phonological limelight, especially in the work of John McCarthy. The dialect has provided illustrations of a number of phonological opacity-related phenomena such as chain-shifting and counterfeeding (Baković 2007; Kirchner 1996; McCarthy 1999; Orgun 1996), Gahawa-Syndrome (McCarthy 1992, 1991), Duke-of-York gambit (McCarthy 2003), and epenthesis-related and metrical opacity (Al-Mozainy 1981; Al-Mozainy et al

²⁰ See Al-Hazmy (1975, 1972) for a general description of the dialect and the Harb tribe.

²¹ Given the place of consonantalism in Arabic, the fact that vowels, rather than consonants, should display extensive alternations might not be paradoxical after all, especially in a conservative dialect such as BHA.

1985; Gordon 2001; McCarthy 2007a, 2007b, 2006, 2003, 1999; Oh 1998). These researchers have used data from BHA to argue for different analyses and tools including the markedness constraints REDUCE (Kirchner) and NO-V-PLACE (McCarthy), the faithfulness constraint MATCH(V) (Orgun), the place feature [pharyngeal] on a par with [labial], [coronal], and [dorsal] (McCarthy), a violation of the otherwise undominated constraint FOOT-FORM (McCarthy), foot-bound stress shift (Al-Mozainy and Al-Mozainy et al), dually-constrained representations (Oh), and Weight-to-Stress analysis (Gordon). McCarthy has also offered an analysis of opacity in BHA within the framework of Sympathy Theory (1999, 2003) and Candidate Chains in OT (2007a, 2007b).

3.2.2 Vowel Epenthesis in BHA

In BHA, vowel epenthesis occurs to break up word-final bi-consonantal clusters with a rising sonority slope.²² In BHA, well-formed word-final clusters are those with a falling sonority slope (see the examples in (11)). In (11), word-final clusters that are not well-formed are marked with *. These are all repaired through vowel epenthesis except those that end in a glide, which are repaired through glide vocalisation. Rising-sonority clusters where vowel epenthesis is required are enclosed in the triple-lined frame in (11). Note also that obstruents (fricatives and stops) are not differentiated for sonority in this dialect.

²² Vowel epenthesis also fixes bi-consonantal clusters violating OCP and breaks up tri-/quadri-consonantal clusters, which arise from morphological concatenation. For the sake of brevity, I will not discuss these here, especially as they will have no bearing on the question of neutralisation that I attempt in this thesis.

(11)

C_# _C#	Glide				
Glide	----- ²³	Liquid			
Liquid	*[marw] 'type of rocks'	-----	Nasal		
Nasal	*[θanj] 'folding'	*[haml] 'carrying'	-----	Fricative	Stop
Fricative	*[bayj] 'aggression'	*[baħr] 'sea'	*[wahn] 'weakness'	[ʕaf] 'baggage'	[baft] 'whitish'
Stop	*[badw] 'Bedouins'	*[ʕadl] 'justice'	*[radm] 'heaps'	[natf] 'plucking'	[ʕabd] 'servant'

Vowel epenthesis is sensitive to the type of prosodo-syntactic boundary²⁴ standing between the word that contains the offending cluster and what follows. For example, for vowel epenthesis to occur, the host word must be either followed by a phrase boundary, as in (12) below, or a consonant-initial tauto-phrasal word, as in (13). Vowel epenthesis does not apply if the following word is vowel-initial, as in (14), unless there is a phrasal boundary separating the two, as in (15).

(12)	*[..CRΦ..]	[..CVRΦ..]
	*[gidr] 'pot'	[gidir]
	*[naħr] 'neck'	[naħar]
	*[ʕadl] 'justice'	[ʕadil]
	*[gidr ħaragna] 'a pot burned us'	[gidir ħaragna]
	*[gidr fa:duhum] 'a pot was of use to them'	[gidir fa:duhum]

(13)	*[..CR#C..]	[..CVR#C..]
	*[gidr nu:ra] 'Nora's pot'	[gidir nu:ra]
	*[naħr samar] 'Samar's neck'	[naħar samar]
	*[ʕadl malik] 'the justice of a king'	[ʕadil malik]

²³ In Arabic, sonorant consonants belonging to the same place-of-articulation class are subject to a special restriction such that they may not co-occur within the same consonantal root (except for nasals). See Greenberg (1950), Gafos (2003b), and McCarthy (1986).

²⁴ See for example Nespor and Vogel (1986), Kainada (2006), and Kaisse (1985) on the role of prosodic boundaries in phonology. On the interaction of phonological processes and morphological levels in a neighbouring Arabic dialect (Makkan), see Abu-Mansour (1992). On the effect of prosodic boundaries on the articulation and acoustics of speech sounds, see for example, Cho (2004, 2002), Fougeron (2001), Tabain and Perrier (2007, 2005) and references therein.

(14)	[..CR#V..]	*[..CVR#V..]
	[gidr um:i] 'my mother's pot'	*[gidir um:i]
	[naħr ami:nah] 'Ameena's neck'	*[naħar ami:nah]
	[ʔadl ima:m] 'the justice of a ruler'	*[ʔadil ima:m]
(15)	*[..CRΦ...]	[..CVRΦ...]
	*[gidr aθaruh] 'a pot it turned out to be'	[gidir aθaruh]
	*[gidr agu:l] 'a pot I'm saying'	[gidir agu:l]

The quality of the vowels inserted within these clusters depends to a large extent on the quality of the preceding lexical vowel within the hosting word. This is particularly true for epenthetic [i] and [u]. When the lexical vowel is /i/, the epenthetic is invariably [i]. Likewise, if the lexical vowel is /u/, the epenthetic is always [u]. This is illustrated in (16) below. When the lexical vowel is /a/, we see consonantal effects on the quality of the epenthetic vowel. Here, the epenthetic is [a] if the preceding consonant is a guttural, as in (17); it is [u] if the following consonant is [r] and the preceding consonant is not a guttural, or if the following consonant is [m] and the preceding consonant is an emphatic, [g], or [k], as in (18). In other words, a [grave] (Jakobson et al 1963) or what Al-Mozainy (1981: 72) calls “backing inducing” component is evident in the environment for [u] (but see (20) below). Elsewhere²⁵, the epenthetic is [i], as in (19). Free variability between [i] and [u] is evident when the consonant preceding the epenthetic is an emphatic and the consonant following is [l], as in (20). Variants like those in (20) are subject to no obvious linguistic or stylistic conditioning factors. In this thesis, I restrict my phonetic analysis to ‘a’ and ‘i’ in words hosting no emphatics (see chapter four). For this reason, the [u]~[i] variation following an emphatic is irrelevant to the present study. See Table 3-3 for a summary of these environments.

(16)	ʃiʔ[i]r	'poetry'
	sit[i]r	'protection'
	riʒ[i]l	'a leg'
	kib[i]r	'conceit'

²⁵ There is general agreement that [i] is the default epenthetic vowel in Arabic (e.g., Kenstowicz 1994: 272; Lombardi 2002).

- | | | |
|------|-----------------------|-------------------|
| | kub[u]r | 'size' |
| | bux[u]l | 'misery' |
| | ʕug[u]m | 'impotency' |
| | ʃuʕ[u]r ^ʕ | 'a fracture' |
| (17) | s ^ʕ ah[a]n | 'plate' |
| | ʃaʕ[a]r | 'hair' |
| | nax[a]l | 'palm tress' |
| | bah[a]m | 'young sheep' |
| | baɣ[a]l | 'mule' |
| (18) | s ^ʕ ab[u]r | 'patience' |
| | gab[u]r | 'grave' |
| | bad[u]r | 'full moon' |
| | baz[u]r | 'kid' |
| | ʕas ^ʕ [u]r | 'afternoon' |
| | hað ^ʕ [u]m | 'digestion' |
| | ʕað ^ʕ [u]m | 'bones' |
| | xas ^ʕ [u]m | 'opponent' |
| | rag[u]m | 'numeral' |
| | ʕag[u]m | 'stumbling block' |
| | lak[u]m | 'fisting' |
| (19) | ħab[i]l | 'ropes' |
| | mat[i]n | 'shoulder' |
| | bat ^ʕ [i]n | 'stomach' |
| | ħað ^ʕ [i]n | 'hugging' |
| | xat[i]m | 'stamp' |
| | ras[i]m | 'drawing' |
| | naʒ[i]m | 'star' |
| | ʕaʒ[i]n | 'kneading' |
| | waz[i]n | 'weight' |
| | ʔak[i]l | 'food' |

ram[i]l 'sand'

- (20) sat^ɿ[u]l~sat^ɿ[i]l 'a bucket'
 rat^ɿ[u]l~rat^ɿ[i]l 'pound: weight unit'
 ʔað^ɿ[u]l~ʔað^ɿ[i]l 'causing harm'
 mas^ɿ[u]l~mas^ɿ[i]l 'serum'

/V ₁ /	[V ₂]	C _∈ {h,ʃ,h,ɣ,x}—R#	C _∉ {h,ʃ,h,ɣ,x}—[r]#	C _∈ {t ^ɿ ,s ^ɿ ,ð ^ɿ ,g,k}—[m]#	C _∈ {t ^ɿ ,s ^ɿ ,ð ^ɿ }—[l]#	elsewhere	
/i/	[i]	Consonantal Environment Irrelevant					
/u/	[u]						
/a/	[a]	Yes					
	[u]	No	Yes				
		No	No	Yes			
	[u]~[i]	No	No	No	Yes		
	[i]	No	No	No	No	Yes	

Note: R= sonorant {l, r, m, n}

Table 3-3: Summary of conditions determining the quality of the vowel ([V₂]) inserted to break up word-final consonant clusters in words of the shape /...VCR#/. Conditions include (1) a vocalic environment (i.e., the quality of the lexical vowel /V₁/) and (2) a consonantal environment involving different types of word-final sonorants and pre-final obstruents. Consonantal environment is only relevant when the lexical vowel is /a/.

3.3 Vowel Epenthesis and the Phonology of Vowel/Zero Neutralisation

3.3.1 Neutralisation through Vowel Epenthesis

For the purposes of this thesis, I assume the following definition of neutralisation:

- (21) Neutralisation obtains when the terms of a contrast at UR are identical at SR.

Applied to vowel epenthesis, this definition refers to a situation where an underlying vowel and an inserted vowel are identical on the surface. Given (21), consider how vowel epenthesis does and does not neutralise an underlying vowel/zero contrast. Suppose that there is a language where the following surface

generalisation holds: every word is C-initial except those beginning underlyingly with /r/, which surface with an initial vowel. If all word-initial Vr sequences arise through epenthesis, there is no vowel/zero contrast, irrespective of whether or not the epenthetic vowel is drawn from the phonemic inventory of the language. This is shown in (22)-(a) below. If there is no underlying contrast, we simply cannot speak of neutralisation, given (21).

Now suppose that there is a language where every 'i' is epenthetic. If lexical vowels of any quality and epenthetic [i] can occur in the same environment (segmental or prosodic), according to (21), there is a contrast but no neutralisation. The same can be said of a language where 'i' is always epenthetic when word-initial but not elsewhere and where other initial vowels are not epenthetic. Here, there is a contrast but no neutralisation. (22)-(b) illustrates both cases. It is only when 'i' can be epenthetic as well as non-epenthetic and can occur in the same environment that we may speak of an underlying vowel/zero contrast being neutralised through vowel epenthesis, given (21). This is shown in (22)-(c).

To bring the examples closer to what we have in BHA, I give (22)-(d) and (22)-(e). Here, there is an underlying vowel/zero contrast. Lexical and epenthetic vowels have the same quality, at least as far as the impressionistic description is concerned. Hence, under the definition in (21), these examples qualify as instances of neutralisation.

(22)

a.	$/\text{CVCVC}/$ [irabal]	$/\text{CVCVC}/$ [labal]	No Contrast
b.	$/\text{CVCVC}/$ [irabal]	$/\text{VCVCVC}/$ [arabal]	Contrast No neutralisation
c.	$/\text{CVCVC}/$ [irabal]	$/\text{VCVCVC}/$ [irabal]	Contrast Neutralisation
d.	$/\text{CVCC}/$ [dibil]	$/\text{CVCVC}/$ [dibil]	Contrast Neutralisation
e.	$/\text{CVCC}/$ [daħal]	$/\text{CVCVC}/$ [daħal]	Contrast Neutralisation

Now consider the further complication in (23) below:

- | | | | |
|------|----|---------|---------|
| (23) | | /dibl/ | /dabil/ |
| | a. | [dibil] | [dibil] |
| | | /daɦl/ | /daɦal/ |
| | b. | [daɦal] | [daɦal] |

In (23)-(a), it looks as though we have two contrasts and two neutralisations: a vowel/zero contrast neutralised through [i]-epenthesis (/..bil/ and /..bl/ both realised as [..bil]) and a high-low vowel contrast neutralised through a process of low-vowel raising (/di../ and /da../ both realised as [di..]).

There is, however, another way of looking at the surface forms in (23)-(a) which involves placing them in the wider context of the phonology of BHA. In general, open-syllable vowels in BHA are subject to reduction whereby /a/ reduces to [i]/[u] while /i/ and /u/ reduce to zero. This is a chain shift effect (see e.g., McCarthy 2003; Kirchner 1996). There are further complications that I describe in §3.3.3 below. As far as (23)-(a) is concerned, though, we can see that /a/ in /dabil/ is reduced to [i] but that the first [i] in [dibil] from /dibl/ is not underlyingly an /a/, nor is it reduced to zero. Put differently, the vowel-reduction process, which applies transparently to /dabil/, interacts opaquely with vowel epenthesis. As long as the distinction in question concerns the presence versus absence of a vowel in a particular phonological context, the situation in (23)-(a) can still qualify as a minimal contrast involving a vowel and a zero. On this view, vowel epenthesis does not preserve the vowel/zero contrast fully in (23)-(a), nor does it neutralise it completely.

The difference between (23)-(a) and (23)-(b) can be summarised as follows. Like many well-studied neutralisations, (23)-(b) involves only one type of opacity, namely clause (c) of Kiparsky's definition—the clause that is mostly about UR-learning. By contrast, (23)-(a) involves two types of opacity: in addition to neutralisation as opacity (clause c), there is underapplication opacity (clause (a)). [i]-epenthesis renders the vowel-reduction generalisation non-surface-true (see chapter one for more on this).

(23)-(a) and (23)-(b) exemplify the two phonological possibilities for vowel epenthesis to neutralise the underlying vowel/zero contrast considered in this thesis. I discuss the opaque interactions between vowel epenthesis and stress in §3.3.2, and between vowel epenthesis and high-vowel-deletion/low-vowel-raising in §3.3.3. These interactions serve to illustrate how vowel/zero contrasts are neutralised differently according to the quality of the epenthetic vowel. Vowel epenthesis phonologically inserts (1) an [a] that sounds and behaves like lexical /a/ and (2) an [i] that sounds like lexical /i/ but whose presence is not recognised by the phonology, beyond repairing an SSP violation.

3.3.2 Stress and Vowel Epenthesis

Like most Arabic dialects, BHA is a quantity sensitive language with final consonant extrametricality (Al-Mozainy 1981; Al-Mozainy et al 1985; Oh 1998). As such, and for the purposes of this brief sketch, I shall treat word-final [..CVCC] and [..CV:C], which have been traditionally described as superheavy syllables, as heavy and mark the extrametrical final C as <C>. I give in (24) a summary of the stress algorithm.

- (24) (Stress is depicted as ´ over the relevant vowel)
- (a) Considering only the last three syllables, stress the rightmost heavy syllable:
- [kita:ti:] ‘schools’; [miká:ti] ‘offices’; [ʔalmaktibá:<t>] ‘the libraries’ ; [máktiba<h>] ‘a library’; [mákta] ‘a desk/an office’; [kitáb<t>] ‘(I/you) wrote’; [kítba<t>] ‘was written (FM.)’
- (b) Otherwise, stress the antepenultimate light syllable:
- [ʔárafa<h>] ‘Arafat’; [gála<m>] ‘a pen’; [ʔalhálaga<h>] ‘the street market’.

Stress assignment is regular and insensitive to the quality of the lexical vowels. As to epenthetic vowels, there is a difference. More specifically, epenthetic [a] behaves just like lexical vowels, whereas [i]-epenthesis renders stress assignment opaque. Unlike the corresponding lexical vowel, epenthetic [i] is metrically invisible. When it occurs in a stress site as per the algorithm in (24), it fails to bear the stress. Furthermore, epenthetic [i] fails to participate in the syllable-counting function for

stress. In contrast, epenthetic [a] is both stressable and metrifiable, hence, metrically indistinguishable from the corresponding lexical vowel.

The stress pattern of ?al-prefixed words that contain epenthetic vowels illustrates the disparate metrifiability of epenthetic [i] and [a]. For example, in (25), stress uniformly falls on the first syllable of the unprefixed words, irrespective of the underlying status and quality of V₂ (lexical /a/ in (25)-(a); lexical /i/ in (25)-(b); epenthetic [a] in (25)-(c); and epenthetic [i]/[u] in (25)-(d).

(25)	(a).	/hakam/	[hákam]	'a judge'
	V ₂ : /a/	/ʕalam/	[ʕálam]	'a flag'
		/malak/	[málak]	'an angel'
	(b).	/mahil/	[míhil]	'place'
	V ₂ : /i/	/malik/	[málik]~[mílik]	'king'
	(c).	/baħr/	[báħar]	'sea'
	V ₂ : [a]	/ð ^ʕ ahr/	[ð ^ʕ áħar]	'back'
		/ʃaʕr/	[ʃáʕar]	'hair'
	(d).	/ħabl/	[hábil]	'rope'
	V ₂ : [i]/[u]	/gidr/	[gídīr]	'pot'
		/ʃīʕr/	[ʃíʕīr]	'poetry'
		/tamr/	[támur]	'dates'
		/ð ^ʕ uhr/	[ð ^ʕ úħur]	'afternoon'
		/ʃuʕr/	[ʃúʕur]	'fracture'

However, when these words are ?al-prefixed, epenthetic [a] is treated just like lexical vowels: it is parsed as projecting a light syllable within the three-syllable window where [ʔal], being the rightmost heavy syllable, receives the stress. This is also the scenario for all lexical vowels in (26). I repeat the unprefixed forms for ease of comparison.

(26)		_____	ʔal- 'the...'
	(a).	[hákam]	[ʔálhakam]
	V ₂ : /a/	[ʔálam]	[ʔáʕalam]
		[málak]	[ʔálmalak]
	(b).	[míhil]	[ʔálmihil]
	V ₂ : /i/	[málik]~[mílik]	[ʔálmalik]~[ʔálmilik]
	(c).	[báhar]	[ʔálbahar]
	V ₂ : [a]	[ð ^ʕ áhar]	[ʔáð ^ʕ ð ^ʕ ahar]
		[ʕáʕar]	[ʔáʕʕaʕar]

In contrast, ʔal-prefixed words which contain epenthetic [i] or [u] are stressed differently: the prefix [ʔal] does not receive the stress; it is parsed as the second rightmost heavy syllable; the underlying /CVCC/ containing the epenthetic vowel is parsed as the rightmost heavy syllable, with the epenthetic vowel disregarded. See (27) below.

(27)		_____	ʔal- 'the...'
	(d).	[hábil]	[ʔalhábil]
	V ₂ : [i]/[u]	[gídír]	[ʔalgídír]
		[ʕíʕír]	[ʔaʕʕíʕír]
		[támur]	[ʔattámur]
		[ð ^ʕ úhur]	[ʔað ^ʕ ð ^ʕ úhur]
		[ʕúʕur]	[ʔaʕʕúʕur]

Further evidence for the non-stressability of epenthetic [i] comes from the single environment where epenthetic vowels “are systematically stressed in all [Arabic] dialects” according to Farwanah (1995: 151). This is when an epenthetic vowel breaks up a quadri-consonantal cluster. However, in BHA, epenthetic [i] appears in this environment unstressed even though it seemingly occupies what should be a stress site by (24). See (28) below.

(28) (CCCC in boldface; relevant epenthetic vowels underlined)

Dialect	Epenthetic in CCCC	Example (Gloss: 'I wrote to her')
Makkan	Stressed	[katab btá llaha]
Egyptian	Stressed	[katab btí lha]
Palestinian	Stressed	[katab btí lha]
Lebanese	Stressed	[katab btí lha]
Syrian	Stressed	[katab btí lla]
Iraqi	Stressed	[kitab bít laha]
BHA	Not Stressed	[kitá btí lha] ²⁶

Epenthetic [a]s, on the other hand, are readily stressable just like lexical vowels when they occur in a stress site as in (29) below. Here epenthetic [a] occurs in the rightmost heavy syllable, *and* it bears the stress. Note that epenthetic [i] does not.

(29)	ʃaʕ[á]rku<m>	Cf. *gid[í]rku<m>
	'your-MS.C.-PL.-hair'	'your-MS.C.-PL.-pot'
	baḥ[á]rna	Cf. *ḥab[í]lna
	'your-MS.C.-PL.-sea'	'our ropes'
	lah[á]mhi<n>	Cf. *sih[í]rhi<n>
	'their-FM.PL.-meat'	'their-FM.PL.-magic'

A less compelling piece of evidence for the stressability of epenthetic [a]s comes from what Blanc (1970) calls Gahawa-Syndrome, whereby an [a] is inserted post-gutturally in words of the form /CaGCVC/ (G=guttural) according to one analysis. In some Bedouin dialects including BHA, vowel epenthesis feeds a rule of low-vowel deletion whose application after epenthesis creates forms of the shape [CGaCVC] (e.g., [ghawah] 'coffee'; [nxalah] 'a palm tree'). Interestingly, the post-guttural vowel, which is epenthetic [a], bears the word-stress. There is, however, a fair amount of controversy regarding the synchronic status of the process, which will take us too far afield to discuss (for more, see Blevins & Garrett 1998; Hall 2003; McCarthy 1991).

Where stress is quality-sensitive in other languages, it is lexical vowels, in most cases, that are reported to display such an effect (see Kenstowicz 1997). Reports of

²⁶ There is a variant, [kitá**bt**laha], heard in the speech of educated speakers. In this word, epenthesis occurs at the suffix boundary, which is not a stress site by (24).

epenthetic vowels of different qualities behaving differently with respect to phonological processes including stress are not common. An example is found in de Lacy (2002a): in Shipibo, epenthetic [a] appears in foot heads while epenthetic [i] appears in foot-nonheads. Similarly, Lloret and Jiménez (2006) report that in Alguerese Catalan epenthetic [a] behaves like lexical vowels in sustaining the voicing of a preceding sibilant ([dazid₃ amistos] 'friendly desire'; [dazid₃ aspesjal] 'special desire'), whereas epenthetic [i] does not ([dazit_f i feu] 'bad desire'). They argue that epenthetic [a] occurs at the lexical level and is in the prosodic word, which is a prominent position. In contrast, epenthetic [i] is postlexical and occurs outside of the prosodic word, which is a weak position. The pattern in BHA is different. Consonantal effects apart, it is the case that epenthetic [a] and [i] serve the same purpose and occur in the same environment. In Shipibo and Alguerese Catalan, it is the prosodic position that seems to 'license' the quality of the epenthetic vowel; in BHA, it is rather the quality of the epenthetic vowel which seems to determine its phonological visibility.

3.3.3 High Vowel Deletion (HVD)/Low Vowel Raising (LVR) and Epenthesis

As mentioned above, in BHA open-syllable /a/s reduce to [i]/[u], whereas open-syllable /i/s and /u/s reduce to zero. In rule-based derivational models, the pattern is decomposed into two different rules: Low Vowel Raising (LVR) and High Vowel Deletion (HVD).²⁷ In (30), I give an illustration using the rule formulation and data reported in Al-Mozainy (1981: 47, 53-54) and Kirchner (1996: 342).

²⁷ In models where HVD and LVR are treated as different rules, a counterfeeding relation must be stipulated so that an [i], underlyingly /a/, is not wrongly deleted by HVD. This is how it works.

	/samiʔ/	/samiʔat/
HVD	-----	samʔat
LVR	simiʔ	-----
	[simiʔ]	[samʔat]

See Kirchner (1996) for an OT account that appeals to the durational scale /a/>/i/,/u/> zero.

(30)

COPYRIGHT MATERIAL

However, an underlying open-syllable /a/ does not raise when the following vowel is /a/ and the intervening consonant²⁸ is a coronal sonorant /l, r, n/ or a guttural²⁹ /h, ʕ, ħ, ɣ, x/.³⁰ This is illustrated below in (31).

(31)	/ʒa\$lam/	[ʒalam]	‘scissors’
	/ɣa\$nam/	[ɣanam]	‘sheep’
	/ʕa\$rab/	[ʕarab]	‘Arab’
	/na\$ham/	[naham]	‘greed’
	/wa\$ʕad/	[waʕad]	‘promised’
	/sa\$hab/	[sahab]	‘dragged’
	/na\$ɣam/	[naɣam]	‘music’
	/s ^ʕ a\$xal/	[s ^ʕ axal]	‘lamb’
	/fa\$taħ/	[fitah]	‘wrote’
	/ʕa\$lim/	[ʕilim]	‘knew’

Here, I treat HVD and LVR as a single vowel-reduction process (see Kirchner 1996). The interaction of HVD/LVR with vowel epenthesis gives rise to an underapplication opacity effect: [i]-epenthesis creates the environment for HVD/LVR, but it does not apply. In traditional derivational models, this is captured by rule-ordering, as illustrated in (32) below.

²⁸ There is also another restriction similar to the above but involving, instead, post-guttural open-syllable /a/s. These do not raise if the following vowel is low. I don’t discuss this here.

²⁹ Gutturals are sometimes classified as sonorants and sometimes as fricatives. This dual characterisation reflects their special phonetics and phonology (see e.g., Halle 1995; Hall 2003). On the phonetics of Arabic gutturals, see Zawaydeh (1999) and Butcher and Ahmad (1987).

³⁰ The restriction above seems to implicate vowel-to-vowel coarticulation and coronal transparency (e.g., Paradis & Prunet 1991). It might also be explained by appealing to the hypothesised propensity of sonorants including gutturals to allow an adjacent vowel to overlap and extend over their gestures (Hall 2003). The sketch above is only meant to give a brief description of LVR. A full explanation of the pattern is beyond the scope of this thesis.

(32)

√	/ʕumr/	/tamr/	/ħabl/	×	/ʕumr/	/tamr/	/ħabl/
HVD/LVR	----	----	----	Epenthesis	ʕumur	tamur	ħabil
Epenthesis	ʕumur	tamur	ħabil	HVD/LVR	ʕmur	tumur	ħibil
	[ʕumur]	[tamur]	[ħabil]		*[ʕmur]	*[tumur]	*[ħibil]

Without rule-ordering, these forms seem to be exceptions to the otherwise regular pattern of HVD/LVR. Here we have surface open-syllable high vowels that do not delete and surface open-syllable low vowels that do not raise. It is as though these lexical vowels did not occur in open syllables.

As to epenthetic [a], the phonology of the dialect does not make it clear whether it would pattern with epenthetic [i] or with lexical vowels regarding HVD/LVR. This state of affairs obtains because within-word epenthetic [a] only occurs following /a/ plus guttural (e.g., /wahm/ > [waham] ‘fantasy’). It is not clear whether the words with epenthetic [a]s do not undergo HVD/LVR because they are treated as having no low vowel in an open syllable, in which case epenthetic [a] is just like epenthetic [i] and [u], or if they are treated like a lexical /a/, in which case the process does not apply anyway. As described above (see (31)), HVD/LVR does not affect low vowels in open syllables if the following vowel is /a/ and the intervening consonant is a guttural. And of course the process does not affect low vowels in closed syllables. There is no ordering argument to make simply by looking at [a]-epenthesis and HVD/LVR. This is illustrated in (33) below, where either ordering of HVD/LVR and epenthesis gives the correct forms. The specific ordering of HVD/LVR preceding vowel epenthesis is posited by phonologists working within the rule-based tradition to account for the opacity that [i]-epenthesis creates. Importantly, [a]-epenthesis does not give evidence to the contrary.

(33)

	/ba\$gar/ ‘cows’	/ka\$tab/ ‘wrote’	/ra\$han/ ‘pledged’	/rahn/ ‘mortgage’
HVD/LVR	bugar	kitab	-----	-----
Epenthesis	-----	-----	-----	rahan
	[bugar]	[kitab]	[rahan]	[rahan]
Epenthesis	-----	-----	-----	rahan
HVD/LVR	bugar	kitab	-----	-----
	[bugar]	[kitab]	[rahan]	[rahan]

3.3.4 Summary

As we can see, the underlying vowel/zero contrast is apparently neutralised through vowel epenthesis in BHA. However, to answer the question of whether the contrast is completely obliterated on the surface or not, a closer look into the phonological visibility of the epenthised sound is necessary. Phonological visibility is limited here to vowel stressability and metrifiability. Impressionistically, both epenthetic [a] and lexical /a/ occurring in comparable phonological environments sound the same and behave the same with respect to phonological processes. In contrast, epenthetic [i] and lexical /i/ too sound the same but are sometimes treated differentially by the phonology. For example, epenthetic [i] but not lexical /i/ creates opacity when it interacts with certain phonological processes. I have looked here at two such processes—stress assignment and vowel reduction.

3.4 Conclusion

This chapter documented the phonology of vowel/zero neutralisation in BHA. It presented a detailed picture of vowel epenthesis, vowel/zero neutralisation, and the phonological interactions that illustrate the disparate activities of epenthetic and lexical vowels in the phonology of BHA. The chapter demonstrates how the underlying vowel/zero contrast in BHA is neutralised differently according to the quality of the inserted vowel. There is no surface distinction between epenthetic [a] and lexical /a/. In contrast, under certain conditions, epenthetic [i] and lexical /i/ remain different on the surface. It is an empirical question whether or not the phonetics of vowel/zero neutralisation correlates with its phonology. I set out to investigate this in the next chapter.

4 The Phonetics of Vowel/Zero Neutralisation in BHA

4.1 Introduction

In this chapter, I explore the phonetics of vowel/zero neutralisation in BHA in terms of the following main questions:

- (34) Question 1: Is vowel/zero neutralisation phonetically complete?
 Question 2: Is the phonetic effect (complete or incomplete
 neutralisation) genuine?

I have previously referred to these questions as ‘the completeness question’ and ‘the genuineness question’, respectively. I shall continue to use these short forms to refer to the questions above throughout the rest of this thesis.

With regard to the completeness question, I present acoustic and perceptual data from native speakers of BHA. The acoustic data come from a simple experimental design with one independent variable—the vowel/zero underlying contrast that vowel epenthesis supposedly neutralises. Given the phonology of vowel/zero neutralisation in BHA, which I described in chapter three, we can see the merit of a phonetic study that is dedicated to this topic. In terms of the completeness question, the phonology of vowel/zero neutralisation in BHA presents a unique case where both phonologically complete and incomplete effects exist. More specifically, the vowel/zero contrast is completely neutralised through [a]-epenthesis but incompletely neutralised through [i]-epenthesis. A phonetic investigation of this neutralisation provides a valuable opportunity to test hypotheses involving the phonetics-phonology relation. The findings of such a

study will also add a new twist to the ongoing debate about the genuineness of phonetically complete and incomplete neutralisation.

The results of the production experiment reveal a curious pattern of dis-correlation between the phonetics and phonology of neutralisation. Epenthetic [a] and lexical /a/, which behave the same in the phonology, are distinct in the phonetics: epenthetic [a] is statistically significantly more intense than lexical /a/. Conversely, epenthetic [i] and lexical /i/, which behave differently in the phonology, are phonetically identical. No less curious, though, is the apparent utilisation of a contrastively inactive acoustic cue for the purposes of preserving a contrast that seems to be completely neutralised in the phonology. Perception-wise, epenthetic [a] and lexical /a/ are discriminated less accurately than are epenthetic [i] and lexical /i/. I speculate on the relevance of these unexpected results to the laboratory tradition in the study of the phonetics of neutralisation.

As I pointed out while reviewing the literature in chapter two, the genuineness of the reported experimental outcomes has been an issue for many researchers. More specifically, incomplete neutralisation has been regarded as an experimental artefact brought about by certain paralinguistic factors such as orthography and lexical and contextual effects. I manipulate these variables in the experimental paradigm employed here. My exploration of these effects allows for a theoretically possible orthography-induced effect of complete neutralisation.

I also attempt a variety of statistical pre-testing procedures to evaluate the claim that complete neutralisation is a statistical artefact. Failing to reject the null hypothesis that neutralisation is phonetically complete can be due to the experiment having low statistical power. However, I claim that rejecting the null hypothesis that neutralisation is complete can be due to an experiment that is statistically too powerful or an analysis that violates certain statistical assumptions. To anticipate my discussion of the issue, I claim that the statistical artefactuality applies more convincingly to incomplete neutralisation than to complete neutralisation, contrary to what is commonly assumed.

I show that the genuineness argument can be self-defeating. Incomplete neutralisation can be seen as an experimental artefact, but so can complete neutralisation. The experimental artefactuality argument can equally deny the

genuineness of incomplete and complete effects. By the same token, the statistical artefactuality argument can deny the genuineness of both effects. There seems to be no principled way to use the genuineness argument to deny one effect and legitimise the other.

More importantly, I document the variability of the vowel/zero neutralisation data in both qualitative and quantitative terms. Qualitatively, I show that the acoustic difference between epenthetic and lexical vowels is phonetically variable both in directionality and in state of existence. Quantitatively, I present a variety of summary statistics and graphs displaying inter-speaker, inter-item, and inter-condition variations that define the magnitude and direction of the observed acoustic differences.

The chapter is organised as follows. In §4.2, I present and analyse experimental data bearing on the completeness question. In particular, I deal with the acoustic side of the question in §4.2.1 and with the perceptual side in §4.2.2. Then in §4.2.3, I evaluate the completeness question in light of the emerging patterns of the relation between the acoustics and perception of vowel/zero neutralisation in BHA. I then take up the genuineness question in §4.3; I deal with the experimental side of the question in §4.3.1 and with the statistical side in §4.3.2. Finally, I sum up the artefactuality argument in §4.3.2 and conclude the discussion on the phonetics of vowel/zero neutralisation in §4.4.

4.2 The Completeness Question

4.2.1 A Production Experiment

4.2.1.1 Purpose

I carried out this experiment to investigate the acoustics of vowel/zero neutralisation with this specific question in mind:

(35) Production Question: Is vowel/zero neutralisation in BHA acoustically complete?

Answering this question will hopefully reveal the acoustic part of the wider picture of the phonetic completeness of vowel/zero neutralisation in BHA.

As in all previous laboratory studies on neutralisation, this thesis tests complete neutralisation as a null hypothesis. Here, neutralisation will be deemed acoustically complete if there are no statistically significant differences between epenthetic and lexical vowels in minimal pairs along the acoustic parameters investigated. I describe the acoustic parameters of this study further below.

In addition, the experiment addresses a subsidiary question relating to literacy as a between-group variable. The experimental data, having been acquired from a sample of literate and illiterate speakers, make testing the effect of literacy possible. Of course, testing the effect of literacy as a between-group variable is one way of investigating the influence of speakers' awareness of orthography on the phonetics of neutralisation. Literate speakers presumably come with a knowledge of how the contrast in question is orthographically represented; illiterate speakers obviously lack that knowledge. Whether or not this difference in literacy correlates with a difference in the acoustics of vowel/zero neutralisation in BHA is a question this production experiment attempts. I revisit the issue of orthography when I deal with the genuineness question in §4.3.

4.2.1.2 Method

4.2.1.2.1 Speakers

Seven³¹ native speakers of BHA participated in this experiment. They all belong to one clan known by the name of Mahaamiid, which is part of the Harb tribe. They are all monolingual³² female relatives of mine. There were five literates aged between 22 and 33 with a median age of 27, and two illiterates aged 55 and 60. The median age of the whole sample is 31. None of the literate speakers had lived outside of their birthplace, Makkah, where all recordings took place; the two illiterates had been living in Makkah for at least the previous 45 years. These seven speakers did not participate in the frequency-estimation task reported below. Nor did they take part in the perception test reported in §4.2.2.

³¹ Ten native speakers of BHA were recruited. Three of them, however, dropped out at the orientation stage.

³² Although educated speakers are formally taught English at school, the subjects in this experiment know very little English.

4.2.1.2.2 Materials

	V2 _{lexical}	V2 _{epenthetic}
1	[laḥam] 'shut tight'	[laḥam] 'meat'
2	[naḥal] 'teased'	[naḥal] 'bees'
3	[naḥar] 'yelled at'	[naḥar] 'river'
4	[gaḥar] 'oppressed'	[gaḥar] 'oppression'
5	[daḥal] 'entered'	[daḥal] 'income'
6	[faḥam] 'was out of breath'	[faḥam] 'char coals'
7	[raḥan] 'pledged'	[raḥan] 'mortgage'
8	[naḥar] 'slaughtered'	[naḥar] 'the act of slaughtering'
9	[ʃaḥar] 'scalded'	[ʃaḥar] 'month'
10	[gidir] 'managed/overpowered'	[gidir] 'pot'
11	[kibir] 'grew'	[kibir] 'conceit'
12	[gabil] 'Gabil (name)'	[gabil] 'before'
13	[ðikir] 'remembered'	[ðikir] 'prayers'
14	[fikir] 'came to realise'	[fikir] 'thinking'

Table 4-1: The minimal-pair stimulus set in the production experiment

The stimulus set in this experiment is composed of fourteen [C₁V₁C₂V₂C₃] minimal pairs contrasting with respect to the underlying status of their V₂, which is either epenthetic or lexical. These pairs appear in Table 4-1. As can be seen in the table, there are nine pairs illustrating category 'a', and five pairs illustrating category 'i'.

The stimulus set is balanced both phonologically and lexically. Phonologically, only words of the form [CV₁CV₂C] stressed on V₁ and containing clusters violating SSP when V₂ is underlyingly a zero were selected; words violating OCP were excluded on the grounds that epenthesis triggered by different constraints might not necessarily be realised the same. So in order not to introduce a possible confounding factor, only one epenthesis trigger was chosen. As to the epenthetic vowel quality, [u] was excluded: an epenthetic [u] is highly marked cross-linguistically compared to [a] and [i] (Lombardi 2002). Moreover, very few minimal pairs contrasting /u/ with zero exist in BHA; settling for near-minimal

pairs was considered unwise as it might increase segment-based variability within the data.

Lexically, members of each pair in the stimulus set are matched for frequency. Recent research on word production suggests that certain low-level acoustic effects are attributable to frequency. For example, Munson and Solomon (2004) report that vowels in low-frequency words have extreme formant values and are longer in duration than vowels in high-frequency words. Similarly, Whalen (1992, 1991) reports that low-frequency words are produced with greater duration than high-frequency words. Likewise, Pluymaekers et al (2005) show that an affix is shorter in duration when the hosting word is a high-frequency word. Furthermore, Yun (2007) suggests that the degree of vowel-to-vowel coarticulation is 'conditioned' by frequency.

In the current study, frequency matching was based on subjective judgments by forty-six native speakers, none of whom participated in the production task. An objective frequency estimate is not possible for a number of reasons. Firstly, there is no electronic database of BHA, which is only a spoken vernacular. The situation in most Arabic-speaking communities is that dialects serve the best part of everyday oral interaction, while Standard Arabic is used for written transactions. Secondly, the Arabic writing system, where short vowels are normally not indicated, results in enormous ambiguity if words are to be electronically extracted from a Standard Arabic database. For example, the orthographic form ذَكَر could stand for /ðikr/ 'prayers', /ðakar/ 'male', /ðakar/ 'mentioned', /ðukir/ 'was mentioned', /ðakkar/ 'reminded', /ðukkir/ 'was reminded', /ðakkar/ 'used the masculine form', and /ðukkir/ 'was used in the masculine form'.

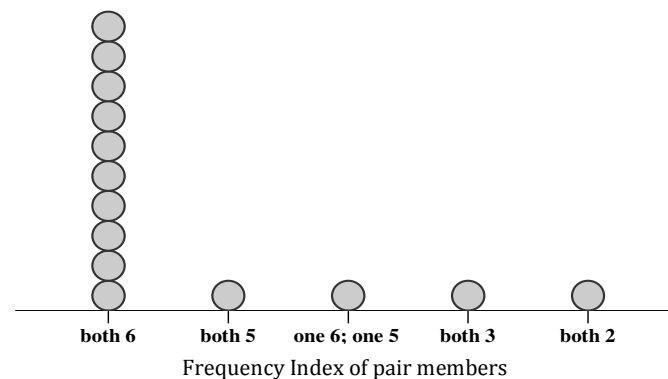
Given this situation in BHA, only a subjective estimate of lexical frequency can be obtained. Importantly, a number of researchers suggest that subjective judgments can be as informative and reliable as objective estimates of frequency (e.g., Carroll 1971; Gernsbacher 1984; Shapiro 1969; Snoeren et al 2006). Interestingly, in work where both are obtained, subjective judgments are found to be more strongly correlated with results of lexical decision tasks (Gordon 1985).

To obtain the ratings used for frequency matching, I first tried Balota et al's (2001) seven-point frequency-of-exposure scale [1= never, 2= once a year, 3= once a

month, 4= once a week, 5= once every two days, 6= once a day, 7= several times a day]. However, it became clear to me that the scale was unnecessarily complicated for a group of participants who belonged to different age groups and had different education levels, and who undertook the task in their homes. It must also be remembered that I had to consult older people who were illiterate. Compare this situation with most frequency-estimation exercises, which use literate subjects and are conducted in classrooms.

Furthermore, the seventeen participants who had to estimate the frequency of the test words on the seven-point scale above consistently ignored 1, 5, and 7 on the scale. That is, for high and low frequency items, participants opted for the less extreme values of 6 and 2, respectively; for medium frequency, they chose the middle values of 3 and 4. Given this initial pattern and the fact that the purpose of the whole exercise was merely to control for frequency, the task was simplified for the rest of my subjects. Now, participants had to rate items on a three-point scale, where the nominal values of 1-2 in Balota et al's scale were collapsed together as [once a year], 5-7 as [once a day], and 3-4 as [once a week to once a month]. The obtained data were then converted to a six-point scale, to accommodate the fractions and middle values that resulted from the calculation of the mean values. The conversion table with the cut-off values is given as Appendix B.

The criterion for selecting the production stimuli was as follows: both members of a pair had to score the same frequency index. In situations where they scored different values, a maximal difference of 1 was accepted. Most selected pairs score 6 on the scale. Figure 4-1 shows the different scores of the pairs in the stimuli.



Note: Each circle represents a pair

Figure 4-1: Stimulus pairs defined by the frequency of their members

A total of fourteen pairs (given in Table 4-1 above) met the inclusion criteria adopted here. The test words were quasi-randomised in such a way that the two members of a minimal pair were not to be found near each other. For example, the list contains /gidir/ and /gidr/ but with many items separating the two. Twenty fillers were appended to the beginning and end of the stimulus list, while at least thirty³³ fillers were inserted between the items.

One obvious concern that remains is the morpho-syntactic status of the stimulus items. Specifically, words with $V2_{\text{epenthetic}}$ are predominantly nouns and those with $V2_{\text{lexical}}$ are almost all verbs. This follows from a general restriction in Arabic morphology where the canonical verb template is CVCVC, while nouns³⁴ can be CVCC.

The phonetics and phonology of individual words are known to be potentially sensitive to lexical category differences. For example, Smith (2001: 61) argues that "nouns show privileged phonological behaviour compared to words of other categories". Similarly, Conwell and Morgan (2007) report that English noun-vowels are significantly longer than the corresponding verb-vowels. Accordingly, the question that suggests itself for the current study is this: if epenthetic vowels turn out to be acoustically different in a systematic way from the corresponding lexical vowels, how can we rule out the possibility that this difference is only a morpho-syntactic effect?

It appears that there are at least two reasons that morpho-syntactic status in Arabic cannot be that influential. Firstly, word-order in Arabic is relatively free. Compare this with English, where nouns occur far more often clause-finally when they serve as an object of a verb phrase and clause-initially when they serve as a subject of a verb phrase. Therefore, English nouns, unlike Arabic nouns or verbs, are naturally expected to be subject to edge lengthening (cf. Klatt 1975). Secondly, the syntactic structure of Arabic, particularly where nouns and verbs are involved, is such that many nouns and many verbs can stand on their own and form one-word sentences. This is in contrast to English, where most ambi-categorial words like 'run' have different structures according to whether they are nouns or verbs. For example,

³³ These within-list fillers were used for other experiments not reported here.

³⁴ There are a few verbs with a $C_1VC_2C_3$ template, but there is no consensus as to whether or not C_3 is a root radical or simply results from C_2 gemination (see McCarthy 1986; Gafos 2003b). There are also CVCVC non-verbs.

the noun version has to occur following a determiner, an article, or a possessive pronoun. A one-word utterance 'run' is not grammatical on a noun-reading (Neil Smith, p.c. 2007). Also, if the verb occurs on its own, it will normally have to be in the imperative. I return to this issue in the discussion section.

4.2.1.2.3 Procedures

4.2.1.2.3.1 Recording

Recordings were made during a trip to Saudi Arabia in the summer of 2007. All subjects were recorded in a quiet room in my family's house. Care was taken to ensure that inside noise was kept to a minimum. The floor, where the room was located, was reserved for the purposes of the experiment. One difficulty was that the air-conditioning systems, a possible source of noise, could not be turned off all the time during the summer. However, the following measures were taken: recording took place in the evening or early morning, when it was less hot; before recording, the room was air-conditioned for some time. A few minutes before the start of a recording session, all air-conditioning systems were turned off.

Although an effort was made to obtain recordings with minimal background noise, outside noise was inevitable in a residential area where all houses are kept air-conditioned round the clock. Another source of outside noise was children playing outside of their homes and cars passing by occasionally. However, it was felt at the time of recording that background noise was not loud enough to undermine the suitability of the obtained recordings for speech analysis.

All recordings were made using a Nagra ARES-M solid-state recorder and a Nagra NM-MICS-II premium quality clip-on mono microphone for ARES-M placed a few inches away from the speaker's mouth. Recordings were digitised at a sampling rate of 22.05KHz. Using Praat (Boersma & Weenink 2008), 588 tokens (14 pairs x 3 repetitions x 7 speakers) were acoustically analysed as per the protocol in §4.2.1.2.3.2.

Target items were orally elicited from each of the seven participants, using orally-presented questions that were designed for this purpose. Participants were instructed to say the response word followed by the word [tara] 'I think', as a two-word clause, with no intervening pause. The inclusion of [tara], which is both syntactically and semantically neutral, was to simplify segmentation before

acoustic analysis. In a pilot, it was noticed that final sonorants and the preceding target vowel were so completely devoiced that any consistent, let alone, reliable segmentation would have been very difficult to maintain. For this reason and since all target words end in a vowel-sonorant sequence, I decided to include the non-target word [tara] pre-pausally.

4.2.1.2.3.2 Acoustic Analysis

4.2.1.2.3.2.1 Acoustic parameters

The obvious target of the acoustic analysis of vowel/zero neutralisation in BHA is V_2 in the $C_1V_1C_2V_2C_3$ words that make up the stimulus set of the experiment. Undoubtedly, looking beyond the relevant obvious target for acoustic cues can add to our understanding of the phonetics of neutralisation. Long-domain distribution of acoustic cues to contrast has repeatedly been described as relevant to speech production and perception (see Alfonso & Baer 1982; Campos-Astorkiza 2007; Hawkins & Nguyen 2003, 2004; Nguyen et al 2009; Wood 1996, among many others). See also Fougeron (2007), Dinnsen (1985), and Fourakis (1984) for a case involving the phonetics of neutralisation. However, given the limitations on this study, I leave this topic for the future.

The study investigates a few well-established acoustic correlates of phonological stress, such as F_0 , intensity, and duration. Focusing on these parameters will increase our potential of finding out more about the acoustics of the phonological visibility of the relevant vowels in BHA. As shown in chapter three, phonological visibility to stress and weight-sensitive processes (i.e., stressability and metrifiability, respectively) illustrates the quality-based distinction in vowel/zero neutralisation in BHA. More specifically, epenthetic [a], like lexical vowels including lexical /a/, is both stressable and metrifiable. In contrast, epenthetic [i], unlike all lexical vowels including /i/, is neither stressable nor metrifiable.

No previous phonetic study of BHA exists. Therefore, my assessment of the relevance of any acoustic parameter has to rely on the available evidence coming from the acoustic studies run on other dialects of Arabic. There is agreement among the studies I reviewed that F_0 , intensity, and duration are correlates of stress in Arabic (de Jong & Zawaydeh 1999; Zuraiq & Sereno 2007). However, a recent study by Bouchhioua (2009) claims that, among the three acoustic

parameters above, only F0 actually cues word-level stress, whereas intensity and duration are associated with phrasal stress. This prosody-based distinction will later guide our interpretation of the results concerning these acoustic parameters.

I have also explored the spectral dimension of the neutralisation data in terms of the first two formant frequencies. To a large extent, the difference between epenthetic [a] and epenthetic [i] is one of quality. There are documentary reports of languages with a quality-sensitive stress pattern (e.g., Kenstowicz 1997; de Lacy 2002a).

4.2.1.2.3.2.2 Segmentation, Labelling, and Acoustic Measurements

As mentioned above, the acoustic analysis in this study targets V_2 in CV_1CV_2C test words. For the purposes of segmentation, V_2 is defined as the temporal interval between two boundary marks, B1 and B2. B1 was hand-inserted at the nearest zero crossing between the designated V_2 and a preceding [x], [h], [h], [b], [d], or [k]. B2 was hand-inserted at the nearest zero crossing between V_2 and a following [l], [r], [m], or [n]. Following Turk et al's (2006: 6) insightful recommendation, I used "more zoomed out spectrogram displays" for locating a boundary region, and "more zoomed in waveform displays" for a more precise placement of the boundary mark within that region.

Generally, as in many acoustic studies (e.g., Alghamdi 2004, 1998; Ham 2001; Lavoie 2001; Piroth & Janker 2004; Ridouane 2007; Warner et al 2004), the main segmentation criterion in determining B1 and B2 was, respectively, the onset and offset of the vowel F2. The onset/offset of the acoustic energy in the F2 region that is characteristic for 'a' and for 'i' on a spectrogram often coincides with an increase/decrease in the amplitude and complexity of the glottal pulsing on a waveform display. Occasionally, I had to consult F3 and its relative intensity (Skarnitzl 2009), in addition to F2 offset, to determine precisely where to insert B2 in V_2 -liquid sequences (see Appendix C for illustrative spectrograms).

The decision to use this set of segmentation criteria, in particular, followed an initial visual inspection of a synchronised wideband spectrogram and waveform display generated by Praat (Boersma & Weenink 2008) for each of the 588 speech files. This form of data pre-viewing shows that in most cases V_2 stands out as a definable spectrographic stretch with characteristic acoustic features even when

following or preceding, respectively, consonants like [h] and [l], which can be notoriously difficult to segment in vocalic contexts (e.g., Olive et al 1993; Skarnitzl 2009; Turk et al 2006). This being the case, the use of the commonly adopted F2-based criterion looks both reasonable and economic. Each labelled interval whose boundaries (B1 and B2) were inserted only once yielded acoustic data for five different parameters. These are mean F0, mean intensity, duration, and F1 and F2 at midpoint.

A custom-written Praat script³⁵ was used to extract the acoustic data along these parameters. For each of the 588 speech files, the script opens a waveform and spectrogram window with label fields at the bottom and another window for checking and manually correcting, where appropriate, glottal pulse markings that appear imposed on a waveform display. These pulse markings were used to calculate mean F0 for each labelled interval. The script also calculates duration, mean intensity, and formants at midpoint using Praat's default Burg algorithm. Finally, the script saves the numerical data as ensemble files, which can then be imported into MS Excel for pre-statistical processing.

4.2.1.2.3.3 Statistical Analysis

I used SPSS (2006) and MS Excel (2003) for pre-test data-processing, including graphing data and calculating an average for each speaker across repetitions and across stimulus items. For more appropriate use of the relevant formal statistical tests, speaker averages, rather than raw data, were submitted to the testing procedures (see Max & Onghena 1999; Raaijmakers 2003; Raaijmakers et al 1999).

To conduct the various significance tests, I used SPSS and PSY (Bird et al 2000). I also used the latter to calculate Cohen's *d* values³⁶ (Cohen 1988) and their 95% Confidence Intervals. Finally, I used MLwiN (Rasbash et al 2009a, 2009b) to produce graphs showing the contribution of the underlying status of *V*₂, subjects, and items to the mismatch between what is predicted on the basis of different combinations of these parameters and what is actually observed. See below for more.

³⁵ The script was written by Yi Xu.

³⁶ Cohen's *d* is a common standardised measure of effect size. Its purpose is to index the degree to which an effect is present (Cohen 1988). It relates the mean difference to some measure of variability of the data in question. In other words, the effect size as measured by Cohen's *d* is not only about the magnitude of a difference but also about how reliably present it is in a dataset (see the discussion in §4.2.1.2.4.1 below).

Among the formal statistical tests I used were 2x2 Mixed Anovas and paired-data t-tests. I used the Mixed Anova procedure to test the following sets of null hypotheses (given below in notation format):

$$(36) \quad \begin{aligned} \text{a-H}_{0.1}: \text{Mean}^{[a]} - /a/ &= 0 \\ \text{i-H}_{0.1}: \text{Mean}^{[i]} - /i/ &= 0 \end{aligned}$$

$$(37) \quad \begin{aligned} \text{a-H}_{0.2}: \text{Mean}^{\text{Literates}} &= \text{Mean}^{\text{Illiterates}} \\ \text{i-H}_{0.2}: \text{Mean}^{\text{Literates}} &= \text{Mean}^{\text{Illiterates}} \end{aligned}$$

Here we have a within-subject factor—the underlying status of V_2 —and a between-group variable—literacy. For each vowel category,³⁷ the Mixed Anova tests the relevant null hypothesis from both sets in (36) and (37) simultaneously. I report F-ratios calculated using Type III Sum of Squares, which is recommended by Milliken and Johnson (2009) and Shaw and Mitchell-Olds (1993) for unbalanced designs.

Another issue that is worth mentioning here concerns multiple testing. Although there is only one within-subject variable (V_2 underlying status), there are five response variables (the five acoustic parameters of the study) for each vowel category. Now testing a single null hypothesis subsuming all five parameters like the ones in (36) and (37) above constitutes multiple testing (see Rietveld & van Hout 2005 for a similar scenario involving linguistic data). In other words, the conventional alpha level of .05, which is set before testing, will exceed the actual alpha level that the global null hypothesis should be evaluated against. Repeatedly testing a single hypothesis can result in spurious significance (Rietveld & van Hout 2005; Godfrey 1985).³⁸ To avoid this, I evaluate the generated test statistics against a Bonferroni-adjusted statistical significance level.³⁹

³⁷ Note that vowel category is not a variable in this study. We are not interested in whether or not 'a' is different from 'i'. Testing a single global null hypothesis that abstracts away from the quality of the vowel could make the interpretation of the statistical outcome shaky (see the discussion section). At the same time, it does not reflect the phonological motivation for testing separate null hypotheses for 'a' and 'i' data.

³⁸ Interestingly, Rothman (1990) and Rex Galbreith (p.c. 2009) warn against exaggerating the importance of correcting for multiple testing.

³⁹ This is calculated by dividing .05 by the number of parameters for each vowel category. For example, in this 5-parameter study, the Bonferroni-adjusted alpha level at 5% is .01.

4.2.1.2.4 Results

4.2.1.2.4.1 Summary Statistics

As part of initial data analysis,⁴⁰ I begin the results section by summarising central tendency, dispersion, and effect size statistics. As is standard practice, the mean (\bar{X} in the equations below) summarises central tendency; the standard deviation (henceforth SD) summarises dispersion. The effect size index I give is Cohen's *d*. Table 4-2 gives the mean and SD values of epenthetic and lexical vowels along each acoustic parameter investigated here. These values are calculated over all the seven speakers and are also broken down by literacy group.

		[a]	/a/		[i]	/i/
F0 (in Hz)	All speakers	210 (15.7)	212 (15.3)		213.3 (14)	215.3 (13.9)
	Only literates	208.8 (15.5)	211.3 (15.6)		210 (11.7)	211.5 (9.8)
	Only illiterates	213.9 (22)	213.7 (20.2)		221.6 (21.2)	224.9 (23)
Intensity (in dB)	All speakers	64 (1.42)	62.7 (1.9)		64.9 (1.3)	64 (2.3)
	Only literates	63.8 (1.6)	62.4 (2.2)		64.6 (1.4)	62.9 (1.7)
	Only illiterates	64.8 (0.63)	63.5 (0.84)		65.9 (0.1)	66.8 (0.1)
Duration (in ms)	All speakers	79.8 (7.6)	77.7 (5)		71.2 (6)	69 (7.5)
	Only literates	80.9 (8.4)	77.2 (5)		72.4 (5.4)	70 (5.7)
	Only illiterates	77 (6.9)	78.9 (6.9)		68.4 (9)	66.9 (13.8)
F1 (in Hz)	All speakers	859.9 (48.2)	862 (38.4)		533.2 (45.7)	531.6 (40.2)
	Only literates	854.9 (52.2)	852.3 (42)		546.7 (37.4)	543.6 (28.8)
	Only illiterates	872.6 (50.9)	886.2 (12.4)		499.6 (61.8)	501.8 (62.7)
F2 (in Hz)	All speakers	1638.5 (160.4)	1636.3 (152.2)		2268 (266.4)	2265 (212.3)
	Only literates	1706.5 (131)	1702.3 (116.9)		2390.6 (182)	2354.6 (169)
	Only illiterates	1468.6 (70.4)	1471.4 (90.4)		1962 (176)	2041.4 (127.6)

Table 4-2: Mean and (SD) values of the five acoustic parameters of the study for both 'a' and 'i' across literates and illiterates (i.e., all speakers), only literates, and only illiterates

⁴⁰ I owe this phrase to Chatfield (1995).

As we can see from the figures in Table 4-2, the mean and SD values of epenthetic and lexical vowels along each of the five acoustic parameters are very close, with exceedingly small differences overall. I will say more about the magnitude and directionality of the mean differences below. But let us focus, for now, on the literacy effect. To get a clearer idea, let us compare the performance of literates and illiterates as individuals. Figure 4-2 and Figure 4-3 graph mean values for each individual speaker as well as for the group as a whole.

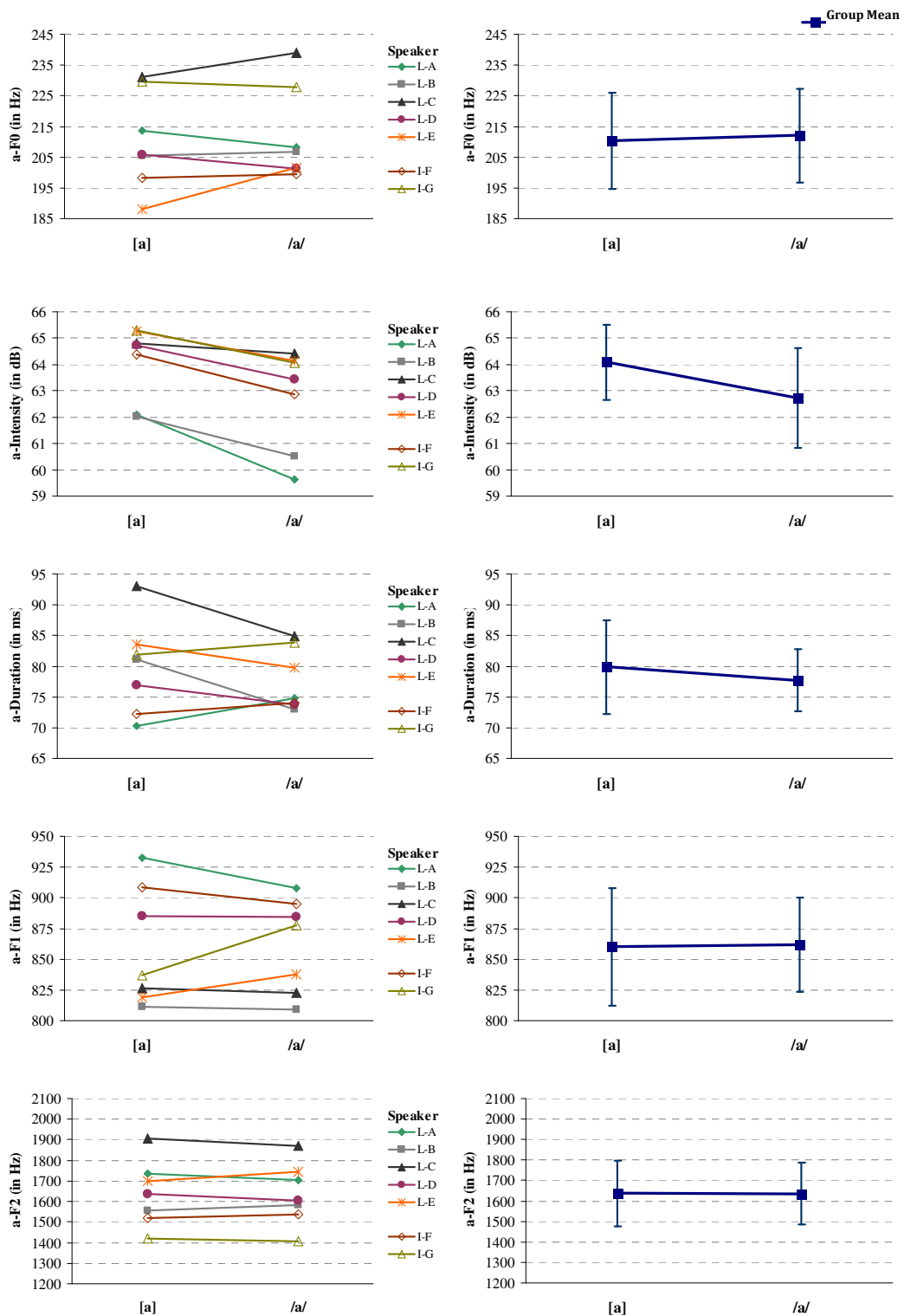


Figure 4-2: Mean values of epenthetic [a] vs lexical /a/ along the five acoustic parameters of the study by individual speakers on the left, and over all speakers on the right $\pm 1SD$. For more clarity, the charts on the left do not display SD values.

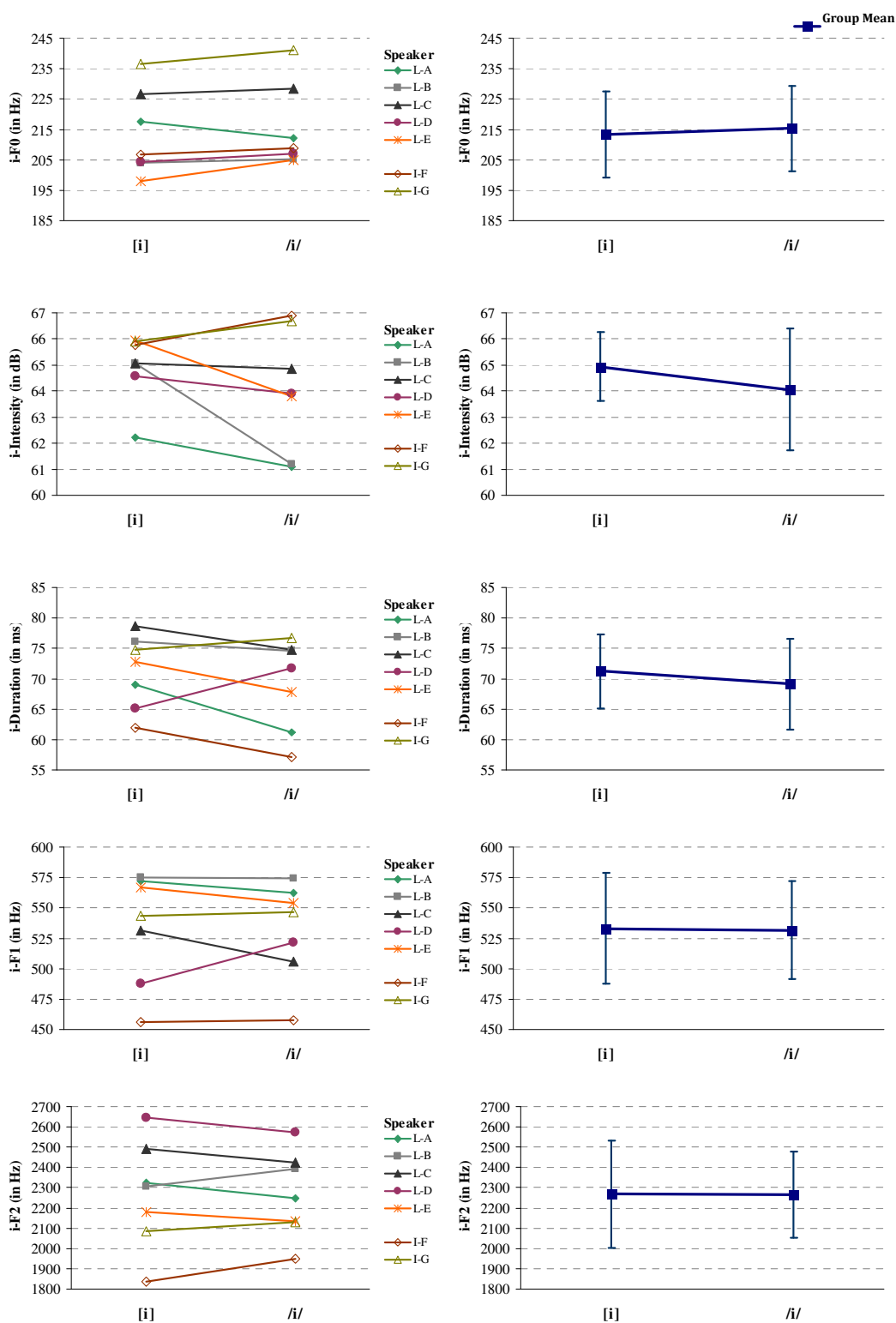


Figure 4-3: Mean values of epenthetic [i] vs lexical /i/ along the five acoustic parameters of the study by individual speakers on the left, and over all speakers on the right $\pm 1SD$. For more clarity, the charts on the left do not display SD values.

The individual-data graphs do not seem to support the proposed grouping of the participants into literates and illiterates.⁴¹ More specifically, the two illiterate speakers do not seem to form a clearly definable and exclusive set whose members share common characteristics that readily distinguish them from the members of the other group. For the purposes of initial analysis, I assume that these characteristics are the mean values and directionality of the acoustic differences between epenthetic and lexical vowels along the parameters of the study.

Looking at the graphs on the left of both Figure 4-2 and Figure 4-3, we observe that the mean-value criterion does not justify a literacy-based grouping: the mean values of the two illiterates are most often further apart from each other than from the values of the literates. This is quite obvious for F0, duration, and F1. For a-F2 and i-F2, it is the case that one illiterate speaker scores a mean value that is far closer to a score by a literate speaker than by the other illiterate speaker.

As to the directionality criterion, it seems that only 'i'-intensity can be taken to suggest a grouping effect along the lines proposed here—literacy-based⁴². Apart from 'i'-intensity, there seems to be no support for treating the speakers as coming from two independent groups. Tukey's nonadditivity procedure which tests for interactions between subjects and conditions (i.e., V₂-UR-Status) failed to find any statistically significant interactions. Furthermore, the significance tests reported in the next section found no statistically significant differences between literates and illiterates along any of the parameters for either vowel category.

In what follows, I make no reference to literacy as a grouping factor in the dataset. Instead, I treat the dataset as paired data from seven participants who, more or less, belong to one group, as far as the data at hand are concerned.

Perhaps among the best summary statistics for paired data are the mean paired difference (\bar{X}_{PD}) and its SD_{PD} . As a preview of the inferential procedures that I

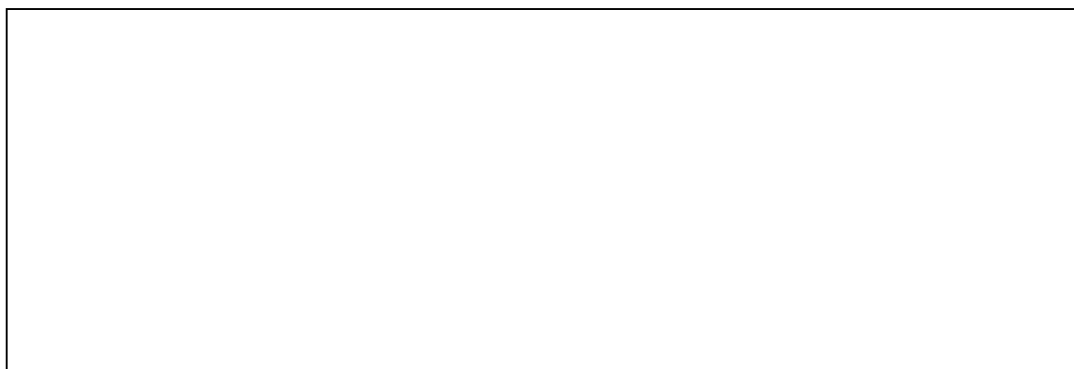
⁴¹ Age-based grouping is as plausible as literacy-based grouping of the sample we have. Recall that age correlates with literacy here: older speakers are illiterate while speakers who are young are literate.

⁴² Again, the effect is also consistent with an age-based grouping.

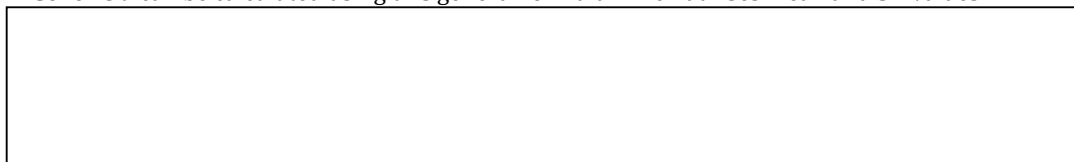
carried out, I report \bar{X}_{PD} , SD_{PD} and relative SD_{PD} (henceforth RSD_{PD}) values. I also report relative mean paired differences ($R\bar{X}_{PD}$).⁴³

Thus, we have two measures of difference size: one absolute (\bar{X}_{PD}), the other relative ($R\bar{X}_{PD}$). We also have absolute and relative measures of variability: SD_{PD} and RSD_{PD} , respectively. A measure relating mean differences to a standardised measure of variability is Cohen's d .⁴⁴ All these descriptive statistics appear in Table 4-3. Figure 4-4 charts d values and their 95% confidence intervals for all the acoustic parameters of the study.

⁴³ These are calculated using the following formulas.



⁴⁴ Cohen's d can be calculated using this general formula which utilises mean and SD values:



However, the simulation study these authors report actually shows that for paired-data designs, using the general mean-SD formula above or the t_c -based equation to derive d gives very similar figures. The package, PSY (2000), which I have used to obtain the d figures I report here, uses the t_c equation above (Bird et al 2000). But I was also able to get the same figures using the general mean-SD formula.

	'a': [a] - /a/					'i': [i] - /i/				
	\bar{X}_{PD}	SD_{PD}	RSD_{PD}	$R\bar{X}_{PD}$	d	\bar{X}_{PD}	SD_{PD}	RSD_{PD}	$R\bar{X}_{PD}$	d
F0	-1.73	6.7	3.9	.008	-.11	-2.02	3.8	1.9	.009	-.14
Intensity	1.36	.61	.45	.02	.811	.87	1.7	2	.01	.46
Duration	2.13	5	2.4	.03	.33	2.08	4.9	2.35	.03	.31
F1	-2.08	21.7	10.4	.002	-.05	1.56	18.6	11.9	.003	.04
F2	2.16	31.6	14.6	.001	.01	3	82	27.32	.001	.01

Table 4-3: Summary statistics of the paired data of the study including central tendency measures (\bar{X}_{PD} : mean paired difference; $R\bar{X}_{PD}$: relative mean paired difference), variability (SD_{PD} : standard deviation of mean paired difference; RSD_{PD} : relative standard deviation of mean paired difference), and effect size (Cohen's d)

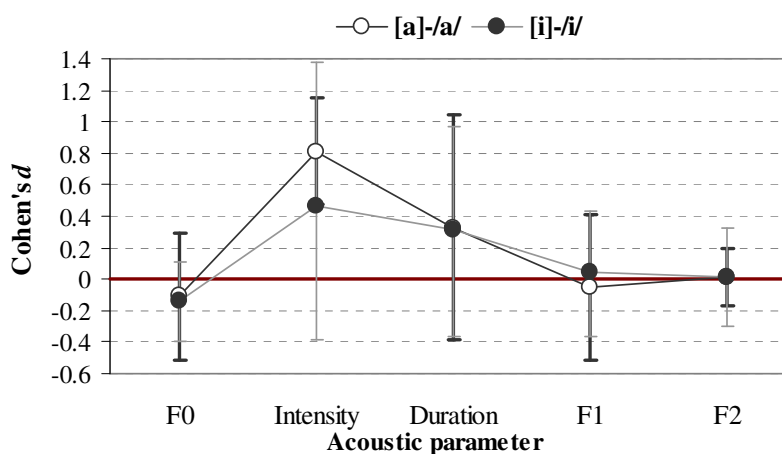


Figure 4-4: Cohen's d and 95% CI values for 'a' and 'i' along the acoustic parameters of the study

The \bar{X}_{PD} figures in the table above show the absolute magnitude and directionality of the acoustic differences between epenthetic and lexical vowels. For both 'a' and 'i', the epenthetic vowel is longer in duration by ≈ 2 ms, lower in pitch by ≈ 2 Hz, and more intense with a larger mean paired difference for 'a' than for 'i'. Mean paired differences along F1 and F2 range from ≈ 1.5 to 3Hz with F2 being higher for epenthetic than for lexical vowels. Also, [i]-F1 is higher than /i/-F1, while [a]-F1 is lower than /a/-F1. But the differences, which are very minute in absolute (\bar{X}_{PD}) and

relative ($R\bar{X}_{PD}$) terms, actually occur against a background of excessive variation. This is what we read from the large SD_{PD} values and the high ratio of SD_{PD} to \bar{X}_{PD} , expressed as RSD_{PD} . These trends are captured by the small d values for most of the parameters.

To get a broader picture, I use these summary measures, ignoring the +/- signs, to construct preliminary ordinal scales of the arithmetic mean, statistical variability, and effect size for each of the two vowel categories 'a' and 'i'. These scales will form the basis of a brief, informal appraisal of the outcome of the inferential tests that I present in the next section.

(38) Ordinal scales of the acoustic parameters of the study established for the summary statistics above

	'a'	'i'
\bar{X}_{PD}	F2>Duration>F1>F0>Intensity	F2>Duration>F0>F1>Intensity
SD_{PD}	F2>F1>F0>Duration>Intensity	F2>F1>Duration>F0>Intensity
RSD_{PD}	F2>F1>F0>Duration>Intensity	F2>F1>Duration>Intensity>F0
$R\bar{X}_{PD}$	Duration>Intensity>F0>F1>F2	Duration>Intensity>F0>F1>F2
Cohen's d	Intensity>Duration>F0>F1>F2	Intensity>Duration>F0>F1>F2

An important issue that we need to explore during initial data analysis is the directionality issue. As described above, the raw mean paired differences display the same directionality for both 'a' and 'i' along all the parameters except F1. However, we can learn a lot more than this by looking into Tukey's mean-difference plots (Tukey 1977; Cleveland 1985). Tukey's mean-difference plots for each of the acoustic parameters for 'a' and 'i' appear in Figure 4-5 and Figure 4-6. In these figures, I have plotted $R\bar{X}_{PD}$ nominator (\bar{X}_{PD}) on the y-axis against its denominator ($(\bar{X}_{epe} + \bar{X}_{lex})/2$) on the x-axis for each individual speaker. For more clarity, I have reproduced these graphs and added to them group values and centroid spikes. Tukey's plots offer a number of advantages for summarising paired datasets, the standard data type in phonetic research. For example, these plots show both the range of mean values and the range of mean paired differences for both individuals and groups. Moreover, the zero-difference reference line

indicates how much and in what direction a paired difference departs from the no-difference zone.

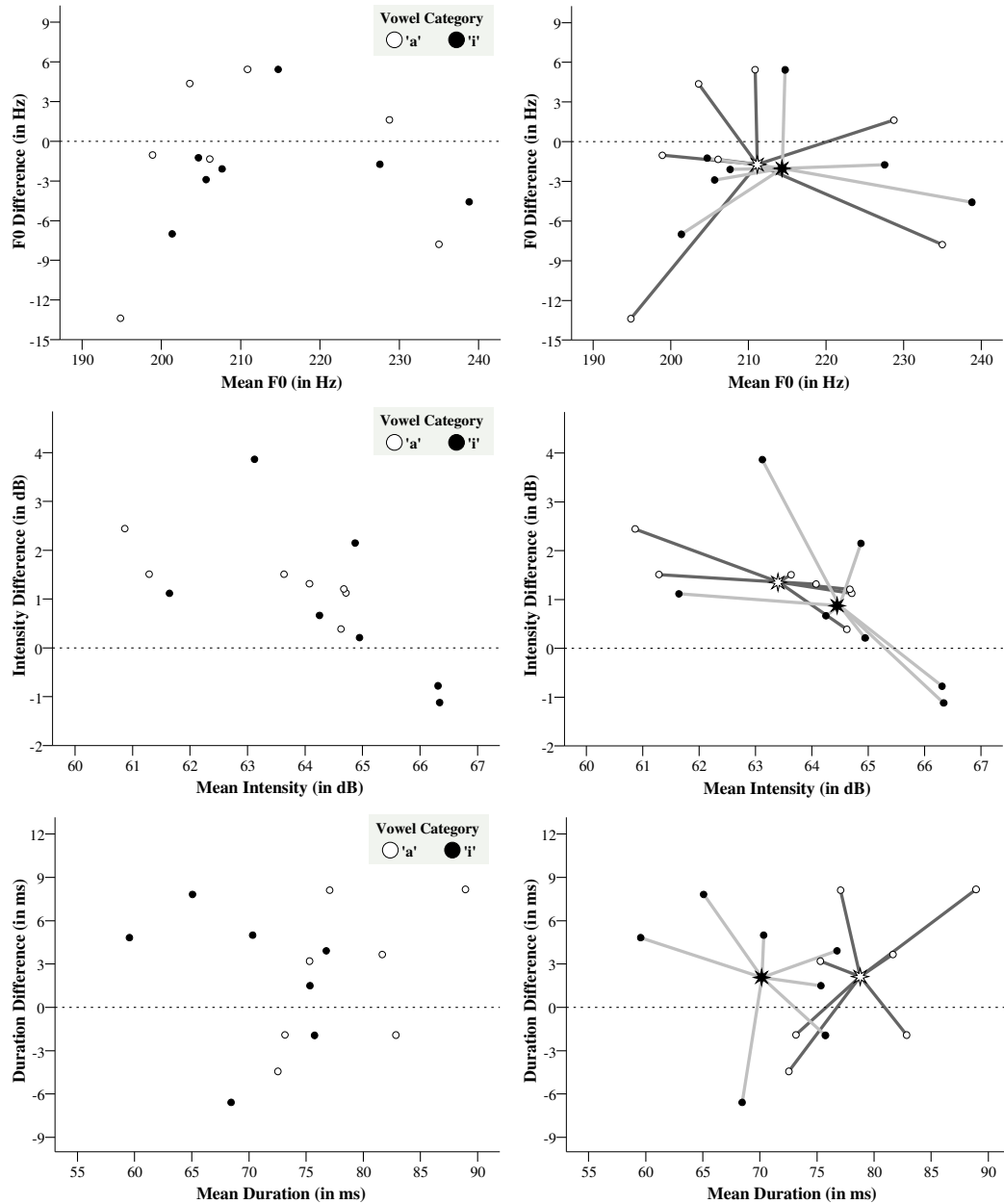


Figure 4-5: Tukey's mean-difference plots for F0, intensity, and duration of 'a' and 'i' data (stars represent group values): $\text{Difference} = \bar{X}_{epe} - \bar{X}_{lex}$; $\text{Mean} = (\bar{X}_{epe} + \bar{X}_{lex})/2$

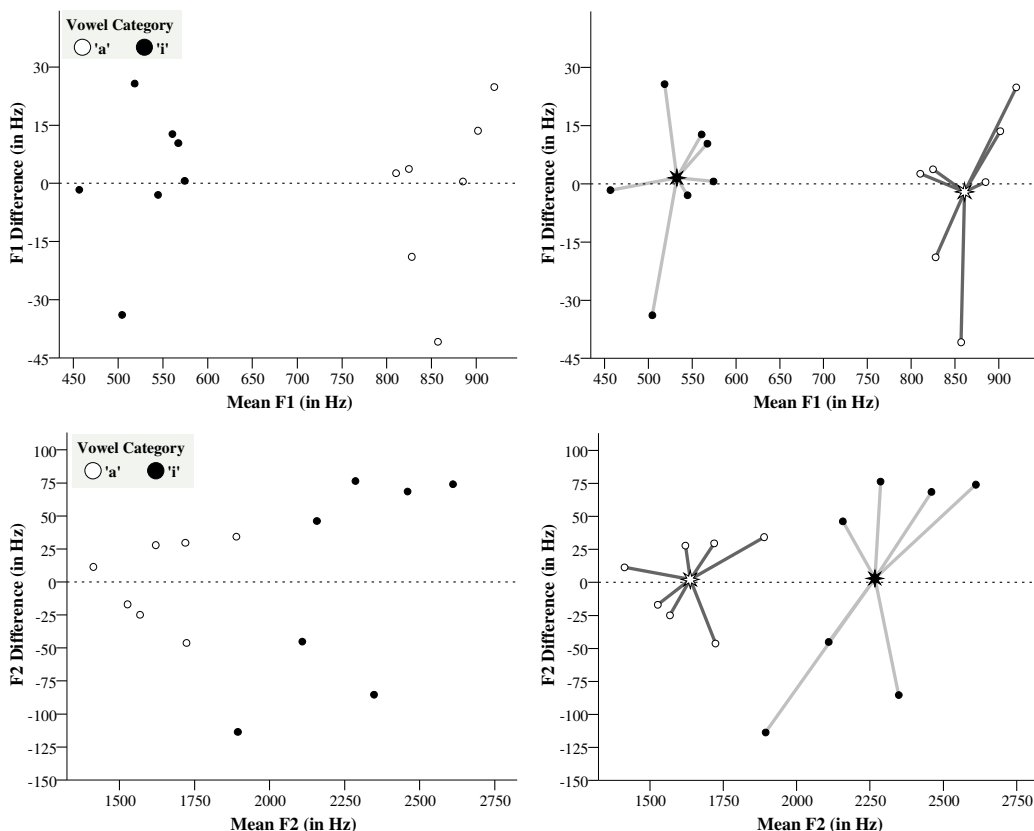


Figure 4-6: Tukey's mean-difference plots for F1 and F2 of 'a' and 'i' data (stars represent group values): $\text{Difference} = \bar{X}_{epe} - \bar{X}_{lex}$; $\text{Mean} = (\bar{X}_{epe} + \bar{X}_{lex})/2$

Looking at the graphs on the right-hand side of Figure 4-5 and Figure 4-6, we observe that the mean paired difference of the group data is closer to the zero-difference reference line than most individual paired differences. This seems to be the case for almost all the parameters for both 'a' and 'i'.

The graphs also show that individual \bar{X}_{PDS} go in opposite directions along all the parameters for both 'a' and 'i' except for a-intensity, where all speakers produce more intense epenthetic [a]s than lexical /a/s. The pattern for i-intensity is very similar, with five speakers out of seven producing more intense [i]s than /i/s. A similar pattern obtains for i-duration, where five speakers produce longer [i]s than /i/s. However, we have the opposite pattern for i-F0, where six speakers have lower-pitched [i]s than /i/s.

Comparing these patterns and others to their corresponding group data, we can see how the distribution and magnitude of positive and negative individual paired

differences define the group mean on the difference axis. That said, we should allow for the possibility that mean differences of similar magnitude but going in opposite directions can cancel each other out, thus levelling the group mean difference towards the zero-difference line and, at the same time, increasing the variance of the data. This is exactly what we observe in the case of i-F2 and, to a lesser extent, a-F2. The group F2 mean difference of [i]-/i/ is 3Hz, while the individual absolute mean differences are far larger than 3Hz. These figures are given in Table 4-4:

Speaker	$ \bar{X}_{PD} $ (in Hz)
L-A	76.4
L-B	85.3
L-C	68.5
L-D	74
L-E	46.2
I-F	113.6
I-G	45.2
Group	3

Table 4-4: i-F2 mean paired differences (in Hz) for each speaker and as a group average in absolute values

This cancelling effect can be demonstrated by a Wilcoxon signed ranks test, which is the non-parametric equivalent of the paired t-test. Roughly speaking, the calculation procedure involves transforming mean differences into ranks, with the smallest difference assigned Rank 1, the next smallest assigned Rank 2, and so on. These ranks inherit the sign (+ or -) of their untransformed values. Next, the ranks are summed and a Z-test statistic is computed, which can be used for null hypothesis testing that involves the median paired difference. Table 4-5 below gives the summary rank statistics of 'a' and 'i' data along the parameters of the study. As can be seen from the table, adding up the positive and negative ranks for i-F2 amounts to zero.

	'a'		'i'	
	$\sum - ranks$	$\sum + ranks$	$\sum - ranks$	$\sum + ranks$
	/a/ < [a]	/a/ > [a]	/i/ < [i]	/i/ > [i]
F0	12	16	6	22
Intensity	28	0	20	8
Duration	20	8	20	8
F1	16	12	16	12
F2	16	12	14	14

Table 4-5: Results of a Wilcoxon signed ranks test for the acoustic parameters of the study for both 'a' and 'i'

Just as the magnitude and direction of the paired differences can vary by speaker, they can also vary across items. Figure 4-7 plots item departures from the group mean. Much of item variation can be attributed to segmental effects. This explains why items ending in the same [...(C)V₂C] sequence are generally closer in their residuals to each other than they are to the other items with a different [...(C)V₂C] sequence. Consider, for example, the three items ending in [...har] and the two items ending in [...ham]. The similarity among the items within these cluster groups is particularly evident in duration, F0, and F1. Intensity residuals of [...ham] items are very similar.

By the same token, the only two i-items sharing a [...CV₂C] sequence—[...kir]—have very similar duration, F0, and F2 residuals in terms of both magnitude and direction. Likewise, [kibir] and [gidir], which share an [...ir] sequence that is preceded by a voiced stop, have similar F0, intensity, F1, and F2 residuals. Of course, there is still a lot of variation. But item variability can be greatly reduced if target sounds are examined within the same or nearly the same immediate phonetic environment. I elaborate on this proposition in chapters six and seven.

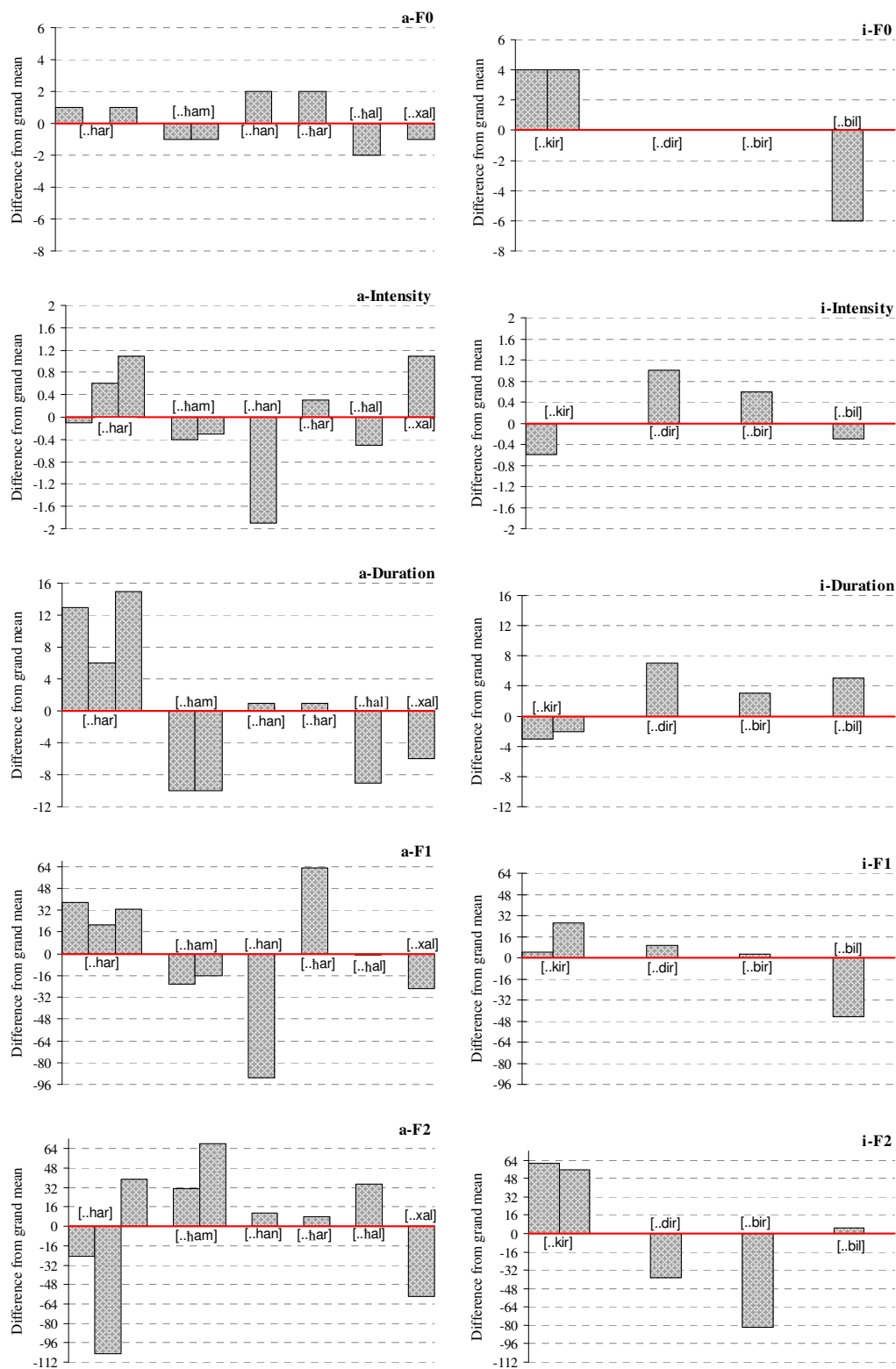


Figure 4-7: Item departures from the group mean values of the acoustic parameters of the study for both 'a' and 'i'

One method to explore and evaluate the contribution of item variation and speaker variation simultaneously is by fitting mixed-effect models, also known as multilevel modelling (see Hox 2002; van der Leeden 1998; for an application of the model to linguistic data, see e.g., Baayen et al 2008; Quené & van den Bergh 2008, 2004). With such models, variation can be partitioned into sources such as speakers, items, and observations comprising every item-speaker combination. I have fitted mixed-effect models to the data following the methodology of Quené and van den Bergh (2008). For technical details, I refer the interested reader to their paper. Here, I only present prediction plots derived from these fitted models. The plots serve to illustrate the sources of variability within the data, including how much variability is due to the underlying status of V_2 in the test items.

The prediction function in the MLwiN package calculates predicted values based on the coefficients and error terms that are fed into the regression equation. The full fitted model has as fixed factors (1) an intercept and (2) a V_2 -UR-Status. The random effects include items, speakers, and observations (i.e., each rendition of an item by each speaker). A prediction plot is generated first using all the components of the fitted model. The prediction calculation then proceeds by excluding one or more of fixed and/or random effect and plotting the resulting values against the observed values. Prediction plots of the incomplete model depict how much it deviates from the full fitted model, to which all fixed and random effects contribute.

The generated plots for the vowel/zero neutralisation data in this study show that the contribution of V_2 -UR-Status is very negligible (see Appendix D for the relevant graphs). The intercept-only model, which is basically the fitted model excluding V_2 -UR-Status, provides as good a fit to the data as the full fitted model that includes V_2 -UR-Status. Of all omissions, omitting the contribution of V_2 -UR-Status causes the least disruption to the model fit. Conversely, the fit seems to lose a lot when prediction is calculated using a regression model that excludes observations.

On the other hand, variation due to subjects is larger than item variation with respect to all acoustic parameters for both 'a' and 'i' except for a-duration, where item variation is larger than subject variation.

4.2.1.2.4.2 Inferential Statistics

Mixed Anovas with V₂-UR-Status (epenthetic vs lexical) as the within-subject factor and literacy (literate vs illiterate) as the between-group variable reveal that for 'a', there is a statistically significant main effect of V₂-UR-Status on intensity [F(1,5)=23.45; p<.006]. V₂-UR-Status has no statistically significant main effect on any of the remaining four acoustic parameters. Nor is there a statistically significant mean difference between literates and illiterates along any of the five parameters investigated here. Moreover, V₂-UR-Status does not interact statistically significantly with literacy for any parameter.

These results provide no support for rejecting the null a-H_{0.2} literacy hypothesis. This agrees with the intuition I outlined in §4.2.1.2.4.1 on the basis of initial data analysis. There seems to be no solid foundation for a literacy-based grouping of the subjects in this study (but see the discussion on NHST in §5.2.2).

The picture for a-H_{0.1} (the neutralisation hypothesis) is less straightforward: evaluating the global hypothesis, we may take a single statistically significant outcome as sufficient evidence to reject this null hypothesis.

As to 'i', Mixed Anovas fail to find any statistically significant results along the five parameters of the study for the neutralisation or literacy hypotheses. Moreover, V₂-UR-Status does not interact statistically significantly with literacy for any parameter. Accordingly, we cannot reject i-H_{0.1} (the neutralisation hypothesis) or i-H_{0.2} (the literacy hypothesis). All Mixed Anovas outcomes appear in Table 4-6.

	'a'						'i'					
	V ₂ -UR-Status		Literacy		V ₂ -UR-Status by Literacy		V ₂ -UR-Status		Literacy		V ₂ -UR-Status by Literacy	
	F(1,5)	p	F(1,5)	p	F(1,5)	p	F(1,5)	p	F(1,5)	p	F(1,5)	p
F0	.14	.73	.07	.79	.22	.66	2.0	.22	1.2	.33	.30	.61
Intensity	23.5	.005*	.53	.50	<.1	.99	.36	.57	6.0	.06	5.4	.07
Duration	.227	.65	.04	.85	2.2	.202	.71	.44	.40	.56	.04	.85
F1	.36	.58	.49	.52	.77	.42	.002	.96	1.8	.24	.10	.76
F2	.002	.96	5.9	.06	.06	.82	.63	.46	6.9	.05	4.5	.08

*statistically significant at the Bonferroni-adjusted alpha at 5%

Table 4-6: Mixed Anovas main effects of V₂ Underlying Status and Literacy and the interactions between these variables for both 'a' and 'i' along the five acoustic parameters of the study

On the basis of these results, I decided to disregard the literacy status of the participants and re-test the V₂-UR-Status effect under separate null hypotheses for ‘a’ and ‘i’, using the same relevant procedures as above. But this time, I submitted the data from all the seven speakers to paired t-tests. The results these tests yield are the same as above for both ‘a’ and ‘i’. In other words, the only statistically significant mean paired difference is an intensity difference separating epenthetic [a] from lexical /a/: [mean difference=1.36dB; SD=.61dB; t(6)=5.87; p< .002]. All other mean paired differences are below statistical significance. All t-tests outcomes appear in Table 4-7.

	‘a’		‘i’	
	t(6)	p	t(6)	p
F0	-.68	.52	-1.4	.21
Intensity	5.87	.001*	1.3	.23
Duration	1.12	.31	1.12	.31
F1	-.25	.81	.22	.83
F2	.18	.86	.10	.93

*statistically significant at the Bonferroni-adjusted alpha at 5%

Table 4-7: Results of two-tailed paired t-tests for ‘a’ and ‘i’ data along the acoustic parameters of the study

Recall from the summary statistics above that a-intensity has a d value of (.811), which is the highest in the dataset. The next highest is i-intensity ($d= .46$). Moreover, variability within a-intensity is the smallest, both in absolute ($SD_{PM}= .61$) and relative ($RSD_{PM}= .45$) terms. But a-intensity mean difference is not the largest, either in absolute ($\bar{X}_{PD}= 1.36$) or in relative ($R\bar{X}_{PD}= .02$) terms. This result is unexpected for the phonetics-phonology relation. It is not clear why a phonetic dimension such as intensity, which plays no contrastive role in vowel systems across the world’s languages, should cue the phonological distinction between epenthetic [a] and lexical /a/. I discuss these results next.

4.2.1.2.5 Discussion

In this section, I use the experimental data described above to (1) construct an account of the acoustics of vowel/zero neutralisation in BHA and (2) discuss the

various implications the results have for the phonetics-phonology relation. I close by highlighting the need for an elaborate experimental paradigm to assess the genuineness of the outcome of our experimentation with the completeness question.

For the purposes of this chapter, neutralisation can be deemed phonetically complete if the five acoustic parameters studied here (F0, intensity, duration, F1, and F2) do not differ statistically significantly according to the underlying status of V₂. With this in mind, we may conclude on the basis of the results above that the vowel/zero contrast is completely neutralised through [i]-epenthesis, but not through [a]-epenthesis. That is, epenthetic [i] and lexical /i/ are acoustically indistinguishable, whereas there is a statistically significant difference in intensity between epenthetic [a] and lexical /a/.

Abstracting away, for the moment, from the identity of the epenthesised vowel, we may conclude that vowel/zero neutralisation in BHA is acoustically both complete and incomplete. This conclusion brings to mind the notion of qualitative variability proposed in chapter two. The phonetics of neutralisation is variable between completeness and incompleteness.

This finding is inconsistent with the various theoretical approaches that have guided researchers' understanding of the phonetics of neutralisation. I reviewed these approaches in chapter two. As a summary, they fall into three main groups with respect to their underlying conception of the phonetics of neutralisation. One group only recognises phonetically complete neutralisation as both genuine and relevant (e.g., Steriade 1999); another group only predicts phonetically incomplete neutralisation (e.g., Ernestus & Baayen 2006); the third group is a combination of the first two, with incomplete neutralisation accepted in certain cases as phonologisation in progress, and complete neutralisation in others as phonologisation accomplished (Barnes 2006).

Now with the kind of results we have, the first two approaches will necessarily reach clashing conclusions: what is acceptable for one is not acceptable for the other. In other words, what the only-complete approach tries to argue for is exactly what the only-incomplete approach tries to argue against. More specifically, the only-complete approach will accept the outcome for 'i', but not the outcome for 'a'.

The exact opposite is true of the only-incomplete approach. This contradiction is actually combined in the either-complete-or-incomplete approach. According to this third approach, vowel/zero contrasts in BHA are only incompletely neutralised through [a]-epenthesis but only completely neutralised through [i]-epenthesis.⁴⁵ It might appear, at a cursory look, that this approach accommodates the experimental data reported so far. However, we should not lose sight of the fact that the model will find itself appealing repeatedly to the genuineness issue to argue away results that are inconsistent with this pattern. The approach needs a more principled explanation of this specific pattern. Unfortunately, the phonology of vowel/zero neutralisation does not rush to its rescue.

The results for both [a]-epenthesis and [i]-epenthesis are not consistent with the predictions in (39) that the phonology of vowel/zero neutralisation seems to support.

- (39) Assuming that the phonetics and phonology of neutralisation correlate closely, epenthetic [a] and lexical /a/, which behave phonologically the same, are predicted to be acoustically and perceptually non-distinguishable, whereas epenthetic [i] and lexical /i/ are predicted to be acoustically and perceptually different, since BHA phonology treats them differently.

What these results suggest is that neutralisation involving [a]-epenthesis, which is phonologically complete, is acoustically incomplete. Conversely, the distinction between epenthetic [i] and lexical /i/, which survives in the phonology, is lost in the acoustic signal (at least along the parameters investigated here)⁴⁶. More generally, the phonetics and phonology of vowel/zero neutralisation in BHA do not mirror each other for either vowel category.

⁴⁵ Another scenario where the approach makes the opposite predictions is also possible. According to this scenario, [a]-epenthesis may only neutralise the vowel/zero contrast completely, whereas [i]-epenthesis may only result in incomplete neutralisation. Note that even in this case, the approach would still have to explain its rejection of the experimental data in a principled way. An appeal to the genuineness issue is not particularly convincing. The experimental design is as naturalistic as possible, and it has already yielded a statistically significant result, if opposite to the approach's predictions.

⁴⁶ It should be noted here that there can only be a limited set of acoustic parameters that can possibly be explored given the usual limitations on time and resources (cf. Jongman 2004; Kopkallı 1993). However, the fact that the very set of parameters investigated in this study has already displayed an epenthetic-lexical difference for 'a' but not for 'i' is interesting in and of itself. This result is worthy of further deliberation. The rest of this chapter explores this finding from a perceptual and statistical perspective.

Instead, they seem to dis-correlate, not just for one vowel type, but for both ‘a’ and ‘i’. The phonetics of neutralisation, as revealed by the statistical analysis of the experimental data, and the phonology of neutralisation, as documented by the pattern of interactions among the phonological processes in BHA that is based on impressionistic data, do not agree on the completeness question. There is a statistically significant acoustic difference between epenthetic [a] and lexical /a/ that the phonology overlooks. At the same time, the phonology treats epenthetic [i] and lexical /i/ differently despite their acoustic non-distinguishability.

This pattern of anti-correlation between the phonetics and phonology of neutralisation compounds the problem for the either-or approach. The pattern seems to defy a principled explanation that does not appeal to the genuineness argument. As I pointed out in chapter two, the genuineness question has frequently been evoked when no theory-based explanation seems plausible. I take up this question in detail in §4.3.

Placing the acoustic completeness of neutralisation, as a null hypothesis, within the context of the phonetics-phonology relation highlights a potential conceptual problem with NHST procedures as currently applied in neutralisation laboratory studies. Rejecting or failing to reject the completeness null hypothesis for both ‘a’ and ‘i’, or rejecting it for ‘a’ while failing to reject it for ‘i’ will have this paradoxical implication for the phonetics-phonology relation: the phonetics and phonology of neutralisation both correlate and dis-correlate depending on the quality of the epenthetic vowel. That is, out of four logically possible scenarios for the phonetics-phonology relation, the NHST-analysed data present us with the weirdest. These scenarios appear in Table 4-8 below.

	‘a’	‘i’	Phonetics-Phonology Relation
Phonology	complete	incomplete	
	complete	incomplete	perfect correlation
Phonetics	complete	complete	partial correlation/non-correlation
	incomplete	incomplete	partial non-correlation/correlation
	incomplete	complete	complete non-correlation (anti-correlation)

Table 4-8: The four logically possible scenarios for the phonetics-phonology relation according to neutralisation effects involving [a]-epenthesis and [i]-epenthesis

Another interesting finding of the experiment involves the directionality issue, which has unduly been neglected in the phonetic research on neutralisation. A laboratory study of the phonetics of neutralisation can miss a lot by failing to take into consideration the directionality of the observed differences. Relying on a group averaged value, which is arithmetically calculated by dividing the sum of all individual data points by the number of data points used in the calculation, can result in an incomplete and potentially misleading generalisation. We have seen in the case of *i*-F2 that individual mean paired differences of similar magnitudes but in opposite directions can cancel each other out in the calculation of the group mean. This is reminiscent of the discrepancy found in the literature between analyses based on group data and those based on data from each individual participant (see e.g., Dinnsen & Charles-Luce 1984; Gouskova & Hall 2009). What data should we use to draw inferences on the phonetics of neutralisation? Is this variability relevant? Is it lawful? Should we not doubt the tools of inference that we have borrowed from psychology and the social sciences? I take up these questions in the next chapters.

Of relevance, though, is a puzzle that descriptive statistics presents regarding the direction of the acoustic differences between epenthetic and lexical vowels in terms of what is phonologically expected and what is phonetically observed. Phonologically, the “ideal epenthetic vowel is one that is least noticeable, i.e., one that is shortest and least sonorous”, to quote Gouskova and Hall (2009: 219). Now the data we have of the phonetics of epenthetic and lexical vowels in BHA seem at odds with this common view in the phonological literature. Here, both epenthetic [a] and [i] are longer in mean duration and greater in mean intensity than the corresponding lexical vowels. Note that the argument still holds even if we choose to drop differences that have failed to reach statistical significance.

Does this take us back to the theme of dis-correlation between the phonetics and phonology of neutralisation? Not necessarily. Recall that the underlying status of V_2 in the stimulus items coincides with a morpho-syntactic distinction. As I have pointed out above, this follows from the morphology of BHA. A possible interpretation of the data in light of morpho-syntactic status is that noun-vowels in BHA are longer and more intense than verb-vowels. Note that this pattern of results seems to be consistent with Conwell and Morgan’s (2007) finding that

noun-vowels in English are longer than the corresponding verb-vowels. However, I observe here that this conclusion is only true for descriptive statistics in my study, which display very negligible mean paired differences. The argument collapses as soon as descriptive statistics gives way to inferential statistics, which has only accorded statistical significance to the intensity difference between epenthetic [a] and lexical /a/. Why should only ‘a’ be longer in nouns than in verbs? Why shouldn’t ‘i’ be as ‘a’ in this regard? Why should vowel quality matter for the morpho-syntax of any language? To salvage this morpho-syntactic account, we obviously need to find a morpho-syntactic argument for a quality-based distinction made in the phonetics of nouns and verbs. Returning to the realm of phonology and phonetics, we know that a quality-based distinction in vowel activity in phonetics and phonology is both demonstrable and real, as this thesis has illustrated.

4.2.2 A Perception Experiment

4.2.2.1 Purpose

The purpose of this experiment is to investigate the discriminability and identifiability of the vowel/zero contrasts that vowel epenthesis in BHA putatively neutralises. The experiment has two perceptual tasks: discrimination and identification. The perceptual data the experiment yields, together with the production data we already have, will form our answer to the main question posed here—the completeness question.

4.2.2.2 Method

4.2.2.2.1 Participants

Twenty-two native speakers of BHA participated in the perception experiment. They were all literate women. None of them participated in the production experiment while all of them took part in the frequency estimation task reported at the beginning of the chapter. None had any known speech or hearing problems. Participants undertook the task in their homes or in my house. They were not paid for their participation.

4.2.2.2.2 Materials

The listening materials in both the identification and discrimination tests were made up of 112 trials of eighteen minimal pairs. Twelve pairs exemplify the

vowel/zero contrast investigated in this thesis; the remaining pairs, which exemplify another vowel-based neutralisation effect, were fillers. The target items were the same as used for the production experiment (see Table 4-1) except that two pairs were excluded. These are [gabil] ‘Gabil’ – [gabil] ‘before’ and [fikir] ‘came to realise’ – [fikir] ‘thinking’. The former pair was excluded because one member is a proper name and the other is an adverb. A proper-name reading might have different accessibility than regular nouns and verbs. In the identification test, responses were limited to ‘noun’ and ‘verb’. There was no way to include an adverb in the test material. The other excluded pair scored the lowest frequency index (2) in the stimulus list. It was excluded for that reason.

The filler items were all real words intended for a different experiment but with basically the same purpose. Each filler pair consists of a noun-member and a verb-member, just like the test pairs in Table 4-1.

All eighteen pairs were extracted from the first repetition of the production-stimuli belonging to one speaker—L-E. By a subjective criterion, this speaker spoke more naturally than the other speakers. By a more objective criterion, however, L-E produced epenthetic-lexical differences whose magnitude in terms of Cohen’s *d* exceeds $|\cdot 3|$ along each of the five acoustic parameters of the study for both ‘a’ and ‘i’.⁴⁷ No other speaker produced differences of equal magnitude along all the five parameters for both vowels. Figure 4-8 graphs speakers’ *d*-values that are above $|\cdot 3|$ for ‘a’ and ‘i’. Table 4-9 summaries mean and SD values of the production data from speaker L-E. Statistical tests run on these data fail to yield statistically significant results for any of the five acoustic parameters (but see §4.3.2 on the validity of statistical tests run on data from a single subject). Raw data along each of the five parameters for each test word in the perception task are given in Appendix E.

⁴⁷ $|\cdot 3|$ was the minimum data-emergent cut-off point that would leave only one speaker when used as a selection criterion.

	[a]	/a/	[i]	/i/
F0 (in Hz)	188 (9)	201.5 (21)	197.8 (17)	204.8 (19)
Intensity (in dB)	65 (1.7)	64 (2.3)	65.9 (1.3)	63.8 (1.5)
Duration (in ms)	83.5 (9.7)	79.8 (6.7)	72.8 (8.8)	67.8 (5)
F1 (in Hz)	818.5 (42.7)	837 (54.8)	567 (29)	554 (26)
F2 (in Hz)	1700 (53)	1746 (45.6)	2180.8 (122)	2048 (87.6)

Table 4-9: Speaker L-E's mean and (SD) values of the acoustic parameters of the study for 'a' and 'i' by V_2 Underlying Status; total number of tokens for each parameter= 84 (14 words X 2 V_2 Underlying Status X 3 repetitions)

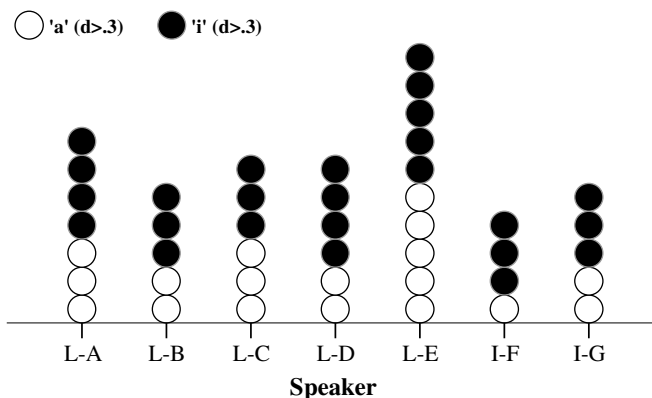


Figure 4-8: Number of acoustic parameters with Cohen's $d > |.3|$ for both 'a' and 'i' by speaker

In preparing the perception stimuli, seventy-six trials of the selected test items and thirty-six trials of the fillers were randomised. Each item has an occurrence rate of $\approx 2.8\%$. The stimuli were arranged into two blocks with an 8-second interval. For two reasons, the amplitude across the words was not equalised. Firstly, intensity is one of the acoustic parameters investigated in the study. Secondly, the NHST-generated results have revealed a statistically significant intensity difference between epenthetic [a] and lexical /a/. In other words, normalising the amplitude, or in fact any other acoustic parameter, might potentially remove a cue to the distinction between lexical and epenthetic vowels.

4.2.2.2.3 Procedures⁴⁸

Participants listened individually to the stimuli over a pair of Audio-Technica ATH-PRO700 professional studio reference headphones connected to a laptop. They had to do a brief mock exercise before taking each test. They received no feedback on their performance. They were allowed to adjust the volume if they wished. In the exercise, participants practised on a different set of words produced by the same speaker as in the perception test. Then they were asked to have a look at the test items, which were printed in Arabic regular orthography, where short vowels were not indicated. The test sheet had the instructions printed in Arabic. Participants also received oral instructions in Arabic at the beginning of the test. Each block of the stimuli appeared on one side of the sheet. Participants were encouraged to take a break after each block. They first did the identification test, which took about 10 minutes, then the discrimination test, which took about 20 minutes. Most participants completed the two tests on the same day, while five did them on two separate days.

In the identification test, participants had to decide for every word they heard whether they thought it was a verb (فعل), in which case V_2 is lexical, or a noun (اسم), in which case V_2 is epenthetic.⁴⁹ They did that by ticking the relevant box on an answer sheet. Inter-stimulus interval (ISI) was 3.5 seconds.⁵⁰

In the same-different discrimination test, participants had to decide whether they thought the two words they heard were the same (نفسه) or different (غيره). They did

⁴⁸ The procedures followed here benefited from post-test feedback from two subjects in a pilot. These subjects did not participate in the experiments reported here.

⁴⁹ Instructing participants to decide directly on V_2 -UR-Status is not possible given that BHA is an unwritten dialect. In Standard Arabic (SA), vowel/zero contrasts can be orthographically represented as follows. The diacritic '*fatha*' [َ] represents /a/; '*kasra*' [ِ] represents /i/; and '*skuun*' [ْ] represents both epenthetic [a] and [i]. However, I have decided against importing this orthographic technique from SA, mainly on the grounds that it might bring along with it unwanted register-transfer effects. See §4.2.1.2.2 for details. More importantly, speakers of BHA are expected to ignore the underlying status of V_2 in an explicit vowel-perception exercise since they would hear the epenthetic V_2 as a vowel with an identifiable timbre. This will increase the number of *fatha* and *kasra* responses, thus biasing the results. After all, the *skuun* is commonly associated with the absence of a vowel within consonant clusters. For this reason, I have made use of the syntactic grouping of the stimulus items, which correlates with the V_2 -UR-Status grouping that is the focus of the experiment (cf. Gouskova & Hall 2009).

⁵⁰ As part of this exercise and the discrimination test, participants had to provide confidence ratings for the responses they gave. I don't report these here.

that by ticking the relevant box on an answer sheet. Within-pair ISI was .8 seconds,⁵¹ while inter-pair ISI was 3.5 seconds.

4.2.2.3 Results

Data from one participant who failed to respond to a number of test items in the identification task were excluded. Excluded, too, were data from two participants who used a single response—‘same’—in the discrimination task.

The overall percent-correct identification is 59% for ‘a’ and 57% for ‘i’. Both results are statistically significantly above chance (50%) by a non-parametric binomial test: [‘a’: $p < .001$; ‘i’: $p < .005$]. Similarly, the overall percent-correct discrimination is 66% ($p < .001$) for ‘a’ and 72% ($p < .001$) for ‘i’.

		‘a’	‘i’			‘a’	‘i’
% Correct Identification	Overall	59%	57%	% Correct Discrimination	Overall	66%	72%
	Noun	66%	60%		Same	87%	89%
	Verb	53%	54%		Different	43%	54%

Table 4-10: Percent correct identification and discrimination of ‘a’ and ‘i’ test words

Do these overall accuracy rates hold equally well for nouns ($V2_{\text{epenthetic}}$) and verbs ($V2_{\text{lexical}}$)? Clearly, we need a break-down of the percent-correct figures above. This is what I give in Table 4-10. According to Table 4-10, participants correctly identified 66% of a-nouns ($V2_{\text{epenthetic}}$) but only 53% of a-verbs ($V2_{\text{lexical}}$). Similarly, they correctly identified 60% of i-nouns ($V2_{\text{epenthetic}}$) but only 54% of i-verbs ($V2_{\text{lexical}}$). The percent-correct identification of a-nouns and i-nouns is statistically significantly above chance by a non-parametric binomial test: [a-nouns: $p < .001$; i-nouns: $p < .001$]. The percent-correct identification of a-verbs and i-verbs, however, is at chance by the same statistical test: [a-verbs: $p = .232$; i-verbs: $p = .301$]. This suggests that the overall accuracy rates in the identification

⁵¹ Admittedly, this ISI is long. However, the purpose of this test is not to tap into listeners’ auditory sensory memory. But rather, it was to give them enough time to process the stimuli using a “labelling strategy” (van Hessen & Schouten 1992). See also Kopkallı (1993). Importantly, Tanner (1961) reports that an ISI less than .8 seconds leads to a decrease in discrimination (in Macmillan & Creelman 2005: 177).

task for both 'a' (59%) and 'i' (57%), which are both statistically significantly above chance, do not hold equally well for both nouns ($V2_{\text{epenthetic}}$) and verbs ($V2_{\text{lexical}}$). Only nouns were identified at an accuracy rate that is above chance; the accuracy rate of verb identification is at chance. See Figure 4-9.

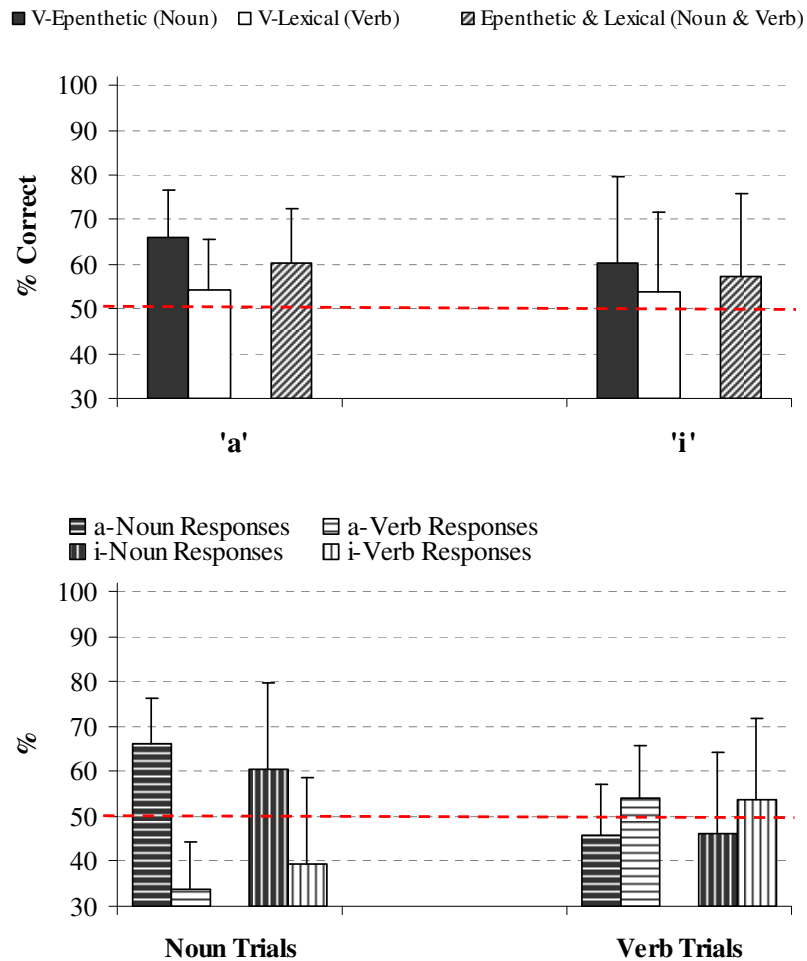


Figure 4-9: Identification of 'a' and 'i' test words according to $V2$ Underlying Status (top) and according to response type (bottom); error bars show 1SD

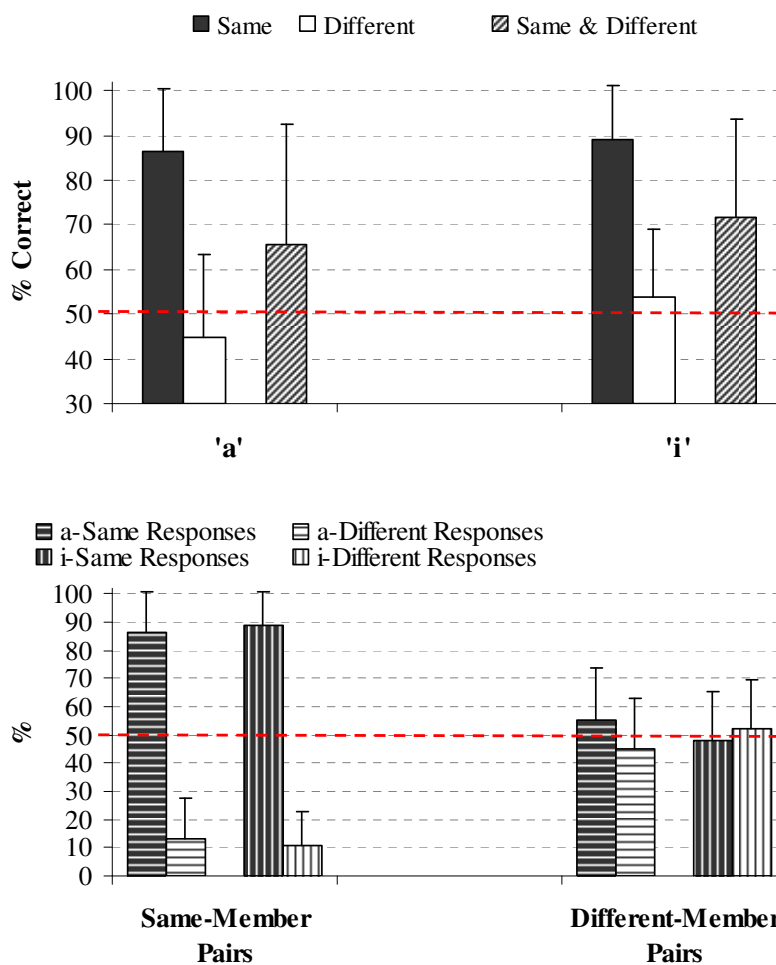


Figure 4-10: Discrimination of 'a' and 'i' test words according to V_2 Underlying Status (top) and according to response type (bottom); error bars show 1SD

The break-down of the discrimination data tells a slightly different story. Table 4-10 above summarises the relevant data. As we can see, 87% of a-same-member pairs and 89% of i-same-member pairs correctly elicited same-responses from the participants. These accuracy rates are statistically significantly above chance for both 'a' ($p < .001$) and 'i' ($p < .001$). Discriminability of i-different-member pairs is at chance ($p = .289$) with a 54% correct different-responses. In contrast, a-different-member pairs only elicited 43% correct different-responses, whereas the wrong same-response rate reached 57%, which is statistically significantly above chance ($p < .002$). These results are graphed in Figure 4-10.

Next, I conducted Repeated Measures Anovas to test null hypotheses involving mean paired differences due to V_2 -UR-Status and to perception tasks. To address

the possibility of a bias (see the discussion section) suggested by the fine-grained analysis above, I adopt Warner et al's (2004) method of looking at the collected responses in terms of one of the two test-alternatives, rather than percent or proportion correct. Accordingly, noun-responses act as the response variable in the identification task, while same-responses act as the response variable in the discrimination task. The choice of these response variables is arbitrary and will not affect the interpretation of the results in any way. Unlike the Warner et al method, however, I have chosen to transform these proportions into Rationalised Arcsine Units (henceforth RAU) (Studebaker 1985; for a recent application to perception data, see Boomershine 2005). Compared with traditional Arcsine-transformations, which are popular in perception research (see e.g., Davidson 2007), RAU-transformations offer the advantage of increased interpretability of the transformed dataset. Specifically, RAUs are very close to the original proportions or percentages. This difference between these types of transformation does not affect their statistical suitability for Anova tests. The RAU figures that I used in the analysis are given in Table 4-11.

Figure 4-11 and Figure 4-12, respectively, display identification RAUs and discrimination RAUs for 'a' and 'i'. For ease of comparison, I have also included the overall correct responses. As suspected, there are far more same-responses than different-responses. Listeners were obviously biased towards same-responses for both 'a' and 'i'. However, a Repeated-Measures Anova with vowel category ('a' vs 'i') and perception task (discrimination vs identification) as fixed factors reveals a statistically main effect of vowel category [$F(1,19)= 5.37$; $p < .04$] and of task [$F(1,19)= 47$; $p < .001$]; the interaction between vowel category and task is not statistically significant [$F(1,19)= .027$; $p = .872$]. Figure 4-13 graphs these effects.

As the graphs show, a-stimuli elicited more noun-responses and more same-responses than did i-stimuli in the identification and discrimination tasks respectively. This suggests that listeners were more willing to respond 'noun' in the identification task and 'same' in the discrimination task for a-stimuli than they were for i-stimuli. In other words, participants showed greater bias with a-stimuli than with i-stimuli. I report bias figures further below.

Interestingly, however, a similar Anova run on the RAUs of all-and-only correct responses failed to find a statistically significant main effect of vowel category

[$F(1,19) = .003$; $p = .95$]. But, there is a statistically significant main effect of task [$F(1,19) = 17.72$; $p < .001$]. The interaction between vowel category and task is statistically significant [$F(1,19) = 5.9$; $p < .03$]. Figure 4-14 graphs these effects. This indicates that [i]-/i/ are discriminated more accurately than are [a]-/a/.

As can be seen from Figure 4-14, [a]-/a/ are identified at a higher accuracy rate than [i]-/i/, whereas [a]-/a/ are discriminated at a lower accuracy rate than [i]-/i/. That is, [i]-/i/ discrimination is better than [a]-/a/ discrimination, while [a]-/a/ identification is better than [i]-/i/ identification. However, paired t-tests reveal that the vowel-based difference in the identification accuracy rate is not statistically significant [$\bar{X}_{PD} = 2.02$; $SD_{PD} = 16.8$; $t(20) = .55$; $p = .59$], while the vowel-based difference in discrimination is [$\bar{X}_{PD} = -5$; $SD_{PD} = 9.4$; $t(19) = -2.33$; $p < .04$].

	Identification RAUs			Discrimination RAUs		
	Noun	Verb	Correct	Same	Different	Correct
'a'	55.6	44.4	58.6	70.5	29.5	65
'i'	53.6	46.4	56.4	66.3	33.7	70

Table 4-11: Mean Rationalised Arcsine Units (RAUs) of the identification and discrimination data for both 'a' and 'i' according to test alternatives and correct responses

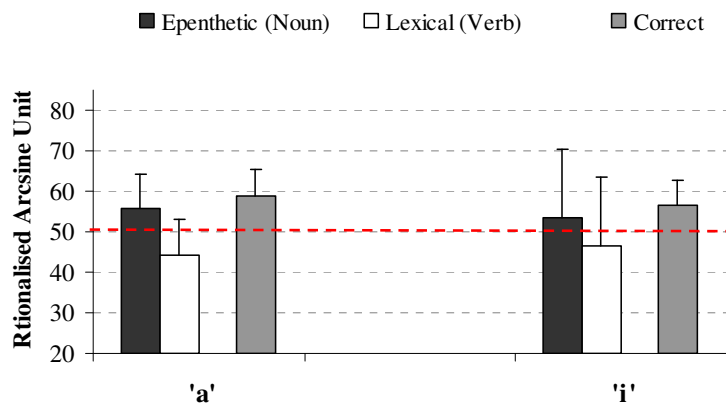


Figure 4-11: Mean Rationalised Arcsine Units (RAUs) of the identification data for both 'a' and 'i' according to test alternatives and correct responses; error bars show 1SD

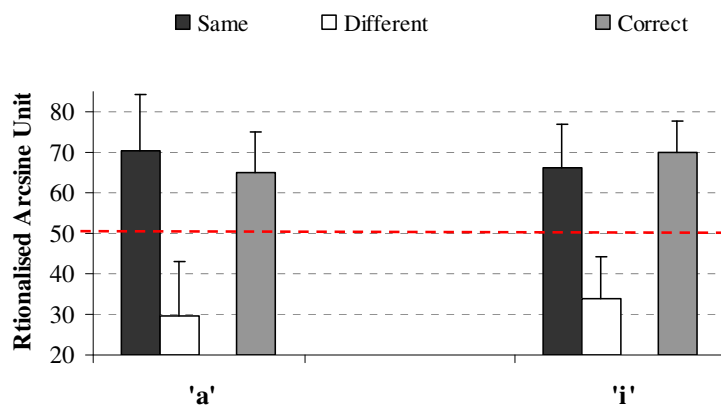


Figure 4-12: Mean Rationalised Arcsine Units (RAUs) of the discrimination data for both 'a' and 'i' according to test alternatives and correct responses; error bars show 1SD

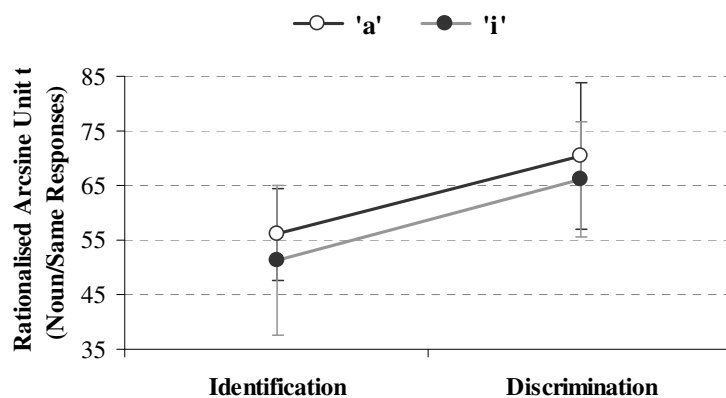


Figure 4-13: Mean Rationalised Arcsine Units (RAUs) of the identification (calculated over noun responses) and discrimination (calculated over same responses) data for both 'a' and 'i'; error bars show $\pm 1SD$

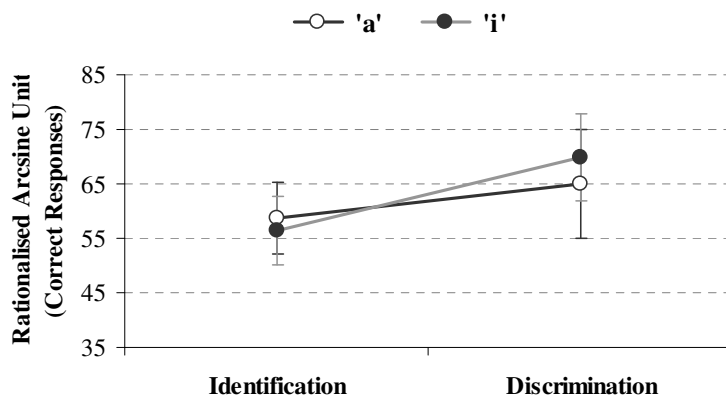


Figure 4-14: Mean Rationalised Arcsine Units (RAUs) of the identification and discrimination data (calculated over correct responses) for both 'a' and 'i'; error bars show $\pm 1SD$

Using techniques from Signal Detection Theory (Green & Swets 1974; Macmillan & Creelman 2005), we can measure respondents' bias and sensitivity. D-prime (henceforth d'), which is a popular index of sensitivity, is roughly defined as the difference between the z-transformed hit proportions and the z-transformed false alarm proportions.⁵² According to the theory, there are four stimulus-response combinations. These are given in Table 4-12 with specific reference to the identification and discrimination tasks in this study. In the identification test, noun-responses to noun-stimuli constitute hits, while noun-responses to verb-stimuli are counted as false alarms. Conversely, verb-responses to verb-stimuli are correct rejections, while verb-responses to noun-stimuli are counted as misses. By the same token, in the discrimination test, different-responses to different-member pairs constitute hits, while different-responses to same-member pairs are false alarms. Conversely, same-responses to same-member pairs are correct rejections, while same-responses to different-member pairs are counted as misses. The direction of the bias and sensitivity measures depends on our initial decision of what constitutes hits and what constitutes correct rejections. As with RAUs, the choice here is arbitrary.

I report here A-prime⁵³ (henceforth A'), a non-parametric analogue of d' (Grier 1971; Johnson 1976; Snodgrass et al 1985; Macmillan & Creelman 2005). See Rallo Fabra (2006) and Wayland and Guion (2003) for a recent application of A' to perception data. A' is said to be more appropriate for proportions nearing 1 and 0, which some participants in the study have. For bias, I used B''_D , the bias measure associated with A' (Donaldson 1992).⁵⁴ A negative B''_D figure represents a liberal bias, whereas a positive B''_D figure indicates a conservative bias. For an unbiased participant, B''_D will be zero. See below for more. A' and B''_D values appear in Figure 4-15 and Figure 4-16, respectively.

⁵² The exact formula for calculating d' differs according to a number of considerations including the type of test paradigm and response strategy (for more see the references in the text above).

⁵³ A' is calculated according to the following formulas, where H= hit rate; FA= false alarm rate (Snodgrass et al 1985: 451):

COPYRIGHT MATERIAL

⁵⁴

COPYRIGHT MATERIAL

		Identification				Discrimination	
		Response				Response	
		Noun	Verb			Different	Same
Stimuli	Noun	hit	miss	Stimuli	Different	hit	miss
	Verb	False alarm	Correct rejection		Same	False alarm	Correct rejection

Table 4-12: Stimulus-response combinations according to Signal Detection Theory as defined for the identification and discrimination tests in this study

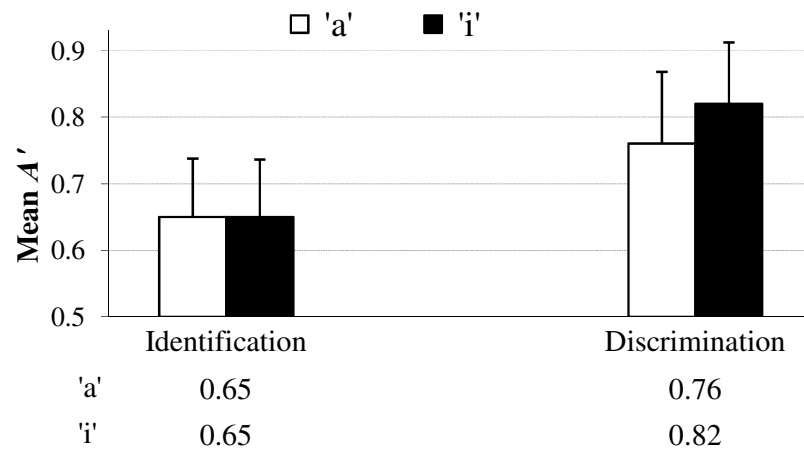


Figure 4-15: Mean A' for 'a' and 'i' by perception task; error bars show 1SD

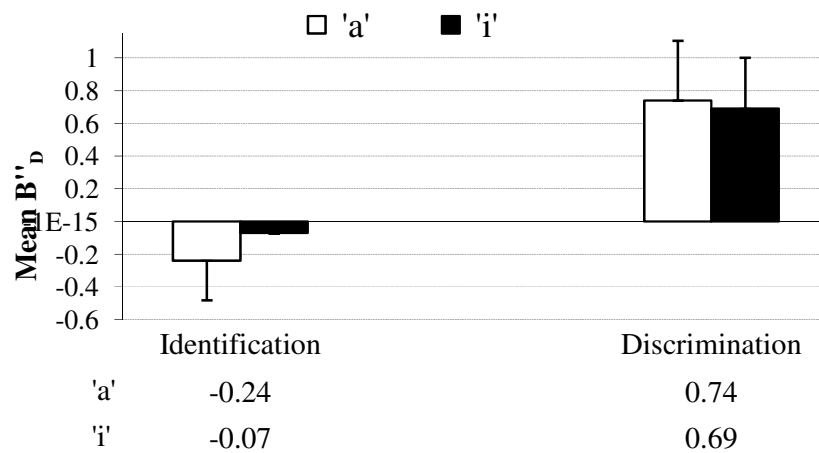


Figure 4-16: Bias for 'a' and 'i' by perception task; error bars show 1SD

Ordering these values from largest to smallest gives these scales:

(40) A' scale: i-discrimination > a-discrimination > i-identification, a-identification

(41) B''_D scale: a-discrimination > i-discrimination > a-identification > i-identification

Taken together, these scales appear to create a paradox: in the discrimination task, participants were both more sensitive and more biased than they were in the identification task. I speculate on this issue in the discussion section below.

4.2.2.4 Discussion

The perception experiment addresses the perceptibility of the vowel/zero contrast in BHA in terms of the discriminability and identifiability of epenthetic and lexical vowels. What we may conclude on the basis of the overall accuracy rates, including all and only correct responses (separately for 'a' and 'i') for each perceptual task, is that the vowel/zero contrast in BHA is perceptible by both the discriminability and identifiability criteria. This perceptibility is statistically significantly above chance with a percent-correct range of 57% – 72%.

However, as Macmillan and Creelman (2005) warn, performance of participants in a perception task can be biased. The observed pattern of responses can sometimes be entirely due to task-dependent or participant-internal criteria or even item- or response-related effects that the analysis overlooks. For example, participants may be inclined to respond in a certain way for a variety of reasons; some items may elicit a particular form of response only too frequently. Obviously, tracing these effects is beyond the scope of this thesis. However, a recommendation that the thesis makes is to include a measure of bias when we report sensitivity indexes such as d' or A' . See below for more on this.

Presenting perceptual data in terms of overall proportion-correct or percent-correct figures can conceal bias and potentially lead to misleading generalisations. Unfortunately, this kind of summary statistics has been the rule rather than the exception in perception studies of neutralisation. The popular claim of an above-chance perceptibility of neutralised contrasts, which has been largely founded on overall percent-correct figures, should not be accepted at face-value (cf. Kopkallı 1993; Jassem & Richter 1989).

As far as the current study is concerned, the break-down of percent correct figures into the test's response-alternatives illustrates how conclusions based solely on collective all-and-only correct responses can be far off the mark.

Our conclusions on the perceptibility of the vowel/zero contrast need to be updated, taking into account the bias we have observed in the responses. Now, what we may conclude from the percent-correct identification of 'a' and 'i' is different for nouns (V2-*epenthetic*) and verbs (V2-*lexical*). The accuracy rate of noun-identification for both 'a' and 'i' is higher than that of verb-identification. It is also significantly statistically above chance. Does this mean that participants were more biased toward noun-responses? Well, according to the bias B''_D figures reported in the Results section, participants were more willing to respond 'noun' than 'verb'. This pattern is true of both 'a' and 'i' although the bias is greater with the former than with the latter. The next logical question is how to explain this bias? It seems to me that the answer lies in the pragmatics of BHA nouns and verbs. A recent experimental study by Sabbagh (2008) shows that, for native speakers of BHA, the noun-reading of ambi-categorial words in pragmatically and syntactically neutral structures is more readily accessible than an equally plausible adjective-reading. Nouns are also more readily accessible than verbs in BHA (Hala Sabbagh, p.c. 2009). A difference in cognitive accessibility of nouns and verbs, if it turns out to be real, will have serious implications for the above-chance identifiability of the neutralised vowel/zero contrast in BHA. It will mean that the 59% and 57% correct identification figures for [a]-/a/ and [i]-/i/, respectively, which are both above chance, are really due to a bias that has nothing to do with the phonetic perceptibility of the contrast. This would effectively imply that the vowel/zero contrast that epenthesis neutralises is not perceptible.

The break-down figures for the discrimination task further confirm the biased pattern of responses. More specifically, the results show that participants were less willing to respond 'different' than 'same'. For 'a', same-member pairs and different-member pairs received a high proportion of same-responses. The percent correct is statistically significantly both above chance for same-member pairs and below chance for different-member pairs. But, the bias, this time, may be attributable to phonetic perceptibility. Recall that data from two participants were excluded. Those two participants gave a single response to all trials—'same'. Compare this

with the data from the one participant who was excluded in the identification task for failing to respond to all test items. The two participants in the discrimination task might have only been reporting what they thought they had heard—no perceptible acoustic differences. Importantly, as Macmillan and Creelman (2005: 218) assert, this response pattern is common for “hard-to-discriminate stimuli”.

More importantly, however, the specific figures we get by breaking down the overall percent correct into the test alternatives actually bring to light a curious effect. Consider for instance *i*-identification. On average, noun-stimuli elicited fewer verb-responses than did verb-stimuli. By the same token, verb-stimuli elicited fewer noun-responses than did noun-stimuli. That is, despite an undeniable noun-bias, participants’ performance seems to be somewhat consistent with a perceptible vowel/zero contrast that epenthesis has putatively neutralised. This pattern is observable in both tasks for both ‘a’ and ‘i’.

Relatedly, Warner et al (2004) suggest that the perceptibility of a neutralised contrast can be revealed by considering the magnitude of a proportion or percent correct not across but within response categories using both correct and incorrect target responses. See also Jongman (2004) and Jassem and Richter (1989).

However, we can learn a lot more by measuring participants’ sensitivity and bias. For example, on the basis of the sensitivity (in terms of A') and bias (in terms of B''_D) figures presented above, I have suggested that participants are both more sensitive and more biased in the discrimination task than in the identification task. This mode of performance, which seems intuitively paradoxical, has a natural explanation, nonetheless.

Sensitivity and bias measure two different properties of participants’ performance. Roughly speaking, sensitivity is about accuracy; bias is about willingness to respond in a certain way (Macmillan & Creelman 2005; Neath & Surprenant 2003). A number of researchers acknowledge that sensitivity and bias are independent of each other (e.g., Donaldson 1992; Macmillan & Creelman 2005, 1990; Snodgrass & Corwin 1988). Certain manipulations of experimental conditions can sometimes effect a change in sensitivity but not in bias. The reverse is also true (see Kamas et al 1996).

The specific pattern of increased sensitivity and bias in the discrimination test as compared to listeners' performance in the identification test is attributable, at least in part, to the nature of the test protocol. Generally, an identification test requires participants to "make comparisons across trials" (Macmillan & Creelman 2005: 184) in order to come up with a labelling criterion. This will naturally delay the formation of the criterion, keeping performance at or just above chance as found in many neutralisation studies. The consequence is decreased sensitivity and decreased bias, as well.

By contrast, in the same-different task, the creation of a response criterion is much easier and quicker. Comparisons occur within rather than across trials (*ibid*). In neutralisation studies, where discrimination is harder than in many other studies, participants are naturally biased towards 'same' responses. The discrimination task reported here creates the most favourable conditions for successfully and easily establishing a criterion of sameness. Recall that the stimuli all come from the first repetitions of the production data from one speaker. Sameness then would simply be identity. Given participants' conservative bias against 'different' and obvious preference for 'same', the following responding strategy seems reasonable: do not respond 'different' unless pair members are *not* identical. However, it seems that participants were successful at establishing what would be 'identical' but not at what would not be. With different-member pairs, participants were merely guessing.

Another equally plausible interpretation which makes more use of the bias account is to say that participants were actually able to detect differences, but they were reluctant to respond 'different'. Recall that identical stimuli elicited a high rate of correct same-responses, while non-identical stimuli elicited what seemed to be a guessing performance despite being heard as non-identical in many more cases than participants were willing to concede. Let us not forget that the group performance on identical trials does not show a ceiling effect. Identical stimuli still elicited between 11% and 13% wrong different-responses. Moreover, the discriminability of the vowel/zero contrast differed according to the quality of the epenthetic vowel: [i]-/i/ were discriminated better than were [a]-/a/. However, the observed pattern of suppressed same-responses for different-member pairs still holds for both vowel categories. All these considerations suggest that

establishing ‘sameness’ is hard, yet possible in many cases. By extension, differences are very small, yet detectable in some cases.

Acknowledging this pattern in my data, I revise the conclusion that the neutralised vowel/zero contrast in BHA is not perceptible. Given what we know so far, we may conclude that this neutralised contrast is still perceptible, irrespective of the quality of the epenthetic vowel. Note that this conclusion does not necessarily imply that [a]–/a/ and [i]–/i/ distinctions are perceived the same. The vowel/zero contrast may still be recoverable for both ‘a’ and ‘i’. Yet the extent of recoverability can be different for these vowel categories.

In summary, the vowel/zero contrast in BHA seems to be only weakly perceptible by both identifiability and discriminability criteria for both ‘a’ and ‘i’ (but see below for a possible counter-argument). At the same time, [i] and /i/ are perceived as more different than are [a] and /a/. These findings are not consistent with the production data described in §4.2.1 above.

4.2.3 General Discussion

The production and perception experiments address this question:

(42) Is vowel/zero neutralisation in BHA phonetically complete?

The production⁵⁵ and perception data in this study suggest that [a]-epenthesis neutralises the vowel/zero contrast in BHA phonetically incompletely: there is an acoustic difference between epenthetic [a] and lexical /a/, which is perceptually detectable. A curious scenario obtains for ‘i’: epenthetic [i] and lexical /i/ are perceptually distinguishable in the absence of a measured acoustic difference. Put differently, ‘i’-data seem to provide a contradictory answer to the completeness question. Specifically, the perceptibility of the contrast is in contradiction with the lack of acoustic differences between the terms of the contrast, epenthetic [i] and lexical /i/. This contradiction would be logically impossible only if we could prove beyond doubt that epenthetic [i] and lexical /i/ are truly acoustically indistinguishable (cf. Dinnsen 1985). But how can we do that? Our

⁵⁵ Considering the perception data in relation to speaker L-E’s production data that have been used in constructing the perception stimuli, we find that the pattern that characterises the ‘i’-data is also true of ‘a’-data. Recall that there are no statistically significant differences for speaker L-E along any of the acoustic parameters of the study for either ‘a’ or ‘i’. But see chapter five for a critique of statistical significance derived from data produced by one speaker.

experimentation can be flawed from data collection to data analysis and interpretation.

That being the case, it may be reasonable to assume that perceptibility necessarily implies an acoustic difference, which can sometimes elude us. This position has actually been adopted by a number of researchers who are concerned with the completeness question. Their argument, however, rests on the premise that there can only be a limited set of acoustic parameters that can possibly be explored given the usual limitations on time and resources (Jongman 2004; Kopkallı 1993). My argument deploys an additional premise: even with the right set of acoustic parameters, using the wrong set of statistical parameters (e.g., measures of variability and central tendency) can lead us astray. In the next chapters, I argue against the use of the arithmetic mean for locating central tendency and the use of SD and RSD for measuring variability. I suggest alternatives that can be utilised within a pro-variability model to address some of the unresolved issues here.

Another reasonable position, if unorthodox, is to assume that it is possible to perceive an acoustic difference that does not really exist. On this view, the contradiction above is merely due to a perceptual illusion. Recall that in the discrimination task, participants labelled 11%–13% of truly identical pair members as ‘different’. This means that, in the identification task, verb-stimuli elicited more verb-responses than did noun-stimuli, not necessarily because there was a V_2 -UR-Status acoustic difference (recall that verbs have $V2_{\text{lexical}}$, while nouns have $V2_{\text{epenthetic}}$), but possibly because some verb stimuli sounded like less than good noun stimuli. Note that this account acknowledges the existence of an acoustic difference, but it denies its relevance to the underlying status of V_2 . That is, all these test items can be accepted as instances of nouns (or of verbs), yet some of them are better instances than others. To many naïve subjects, the two-choice design creates a presumption of relevance of both choices. Just as it happens that BHA speakers are biased towards picking the noun-reading more often, they will treat all better instances as nouns. At the same time, they will treat many less good instances as verbs. A similar argument applies to the discrimination data. Importantly, this account seems to suggest that both [a]-epenthesis and [i]-epenthesis neutralise the vowel/zero contrast completely in the phonetics by the perceptibility criterion. What remains to be explained is the quality-based

distinction in the discrimination data. Recall that [i]-/i/ were discriminated better than were [a]-/a/. We need to explore the perception data against a better-informed account of the acoustics of neutralisation in BHA. I offer to do that in chapters six and seven. Meanwhile, I summarise in Table 4-13 the picture that has emerged so far.

	Phonologically distinct	Acoustically distinct	Perceptually discriminable
/a/ vs [a]	No	Yes (in intensity)	Possibly not
/i/ vs [i]	Yes	No	Possibly yes

Table 4-13: Summary of results for the phonology, acoustics, and perceptibility of the BHA neutralisation data analysed in the study

Clearly, we shall need to undertake further exploration to decide upon the phonetics of neutralisation. However, what we can conclude from the arguments above is that establishing conclusively the phonetic completeness or incompleteness of neutralisation can be very difficult. There is gradience, and there is variability. The question that suggests itself is how relevant and lawful that gradience and variability are? This should naturally bring to mind the genuineness question, which the thesis has posed. I explore this question next.

4.3 The Genuineness Question

I approach the genuineness question from two perspectives: experimental design and statistical treatment. These are the two areas that have inspired and sustained much of the concern and scepticism surrounding the genuineness of the reported findings in the literature. In §4.3.1, I describe an experimental paradigm that addresses the methodological shortcomings that have been identified in the literature. In §4.3.2, I subject neutralisation data from Turkish and Polish to a variety of statistical treatments replicating what has been reported in the literature.

4.3.1 Experimental Artefactuality

4.3.1.1 Rationale for the Experimental Paradigm

The experimental set-up employed by previous studies investigating neutralisation has always been an object of criticism (Baumann 1995; Manaster Ramer 1996a, 1996b; Mascaró 1987) and a subject of further experimentation (Charles-Luce 1993; Jassem & Richter 1989; Piroth & Janker 2004; Port & Crawford 1989; Warner et al 2006, 2004). Whenever the experimental design is blamed, the argument is essentially the following: certain confounds can induce incomplete neutralisation; others can push for complete neutralisation. The possibility of an experimental artefactuality of the reported effect seems particularly strengthened when different studies yield different results even when they are investigating the same neutralisation pattern (see e.g., Fourakis & Iverson (1984) and Port & O'Dell (1985) for *German*; Jassem & Richter (1989) and Slowiaczek & Dinnsen (1985) for *Polish*; Baumann (1995) and Warner et al (2004) for *Dutch*). Given this situation, a number of researchers and reviewers seem more inclined to accept one effect but dismiss the opposite as being simply an artefact brought about by the details of the experiment, including subjects, test materials, and experimental procedures.

Most studies have the following design: subjects are first familiarised with the test materials, which contain minimal or (where not possible) near-minimal pairs and fillers. The subjects, who usually come from different dialectal backgrounds and speak other languages in addition to their native language, read out the test words in a frame sentence. Most of these studies report incomplete neutralisation. Unfortunately, within such an experimental design, there can be a number of factors responsible for whatever effect reported. At least, some of these factors have not been shown to be neutral with respect to contrast realisation and/or neutralisation to date. On the contrary, some have actually been shown to have the potential of inducing incomplete neutralisation.

One such factor concerns the subjects themselves. Lack of homogeneity of the subjects can, if ignored, be a source of variability (cf. Jassem & Richter 1989). This is particularly evident in neutralisation studies that have repeatedly reported a lot of inter-speaker variation and a mismatch between group performance and individual performance. For example, Dmitrieva et al (2010) found that subjects

with some knowledge of a foreign language can produce an incomplete-neutralisation effect greater than or unseen in monolinguals' pronunciation.

In the majority of the previous studies, test materials were presented to subjects in writing. In those cases where the relevant contrast is represented orthographically, most studies report incomplete neutralisation. Conversely, in studies where contrasts are not represented orthographically, neutralisation is mostly found to be complete. Interestingly, Warner et al (2006, 2004) show that an orthographic difference reflecting no underlying contrast can on its own induce speakers to produce a 'sub-phonemic effect'. That incomplete neutralisation is only a spelling pronunciation, though repeatedly claimed, seems imprecise; a spelling pronunciation can more plausibly result in complete neutralisation if the relevant contrast is not represented orthographically (see below for a discussion). Where contrasts are represented orthographically, incomplete neutralisation can be attributed to orthography, underlying representations, surface alternations, etc. Contrasts not represented orthographically but whose neutralisation is found to be complete share with the former all but one factor—an orthographic encoding.

Another factor whose influence has not been fully appreciated concerns the stimuli in terms of presentation and composition. If we review the experimental literature on neutralisation with this in mind, a striking pattern emerges: the majority of the studies whose test material is composed of minimal pairs report incomplete neutralisation. However, when the stimuli contain minimal pairs embedded in sentences with disambiguating clues (e.g., Charles-Luce 1993), or only nonsense minimal pairs (e.g., Herrick 2004), or no minimal pairs at all (e.g., Kopkallı 1993), little or no incomplete neutralisation is reported. This happens despite orthographic representations of the relevant contrasts. Table 4-14 summarises a survey of 25 studies⁵⁶ highlighting the correlation between the absence of minimal pairs in the stimuli and complete neutralisation (see Appendix F for references).⁵⁷

⁵⁶ The inclusion criteria I adopted are as follows. Studies mixing minimal pairs with non-minimal pairs or with nonsense minimal pairs in the test stimuli were excluded. Excluded too were studies exclusively using nonsense minimal pairs. With regard to orthography, only studies providing the orthographic form of the test words as part of the experiment procedures were included. This has added to the reliability and simplicity of the classification decision.

⁵⁷ In the next section, I discuss the statistical validity of the neutralisation findings in the literature.

	Orthography: Contrast represented orthographically	Stimuli: Minimal pairs present	Number of studies reporting <u>neutralisation to be</u>	
			(a): <u>complete</u>	(b): <u>incomplete</u>
1	Yes	Yes	1	16
2	Yes	No	3	1
3	No	Yes	1	1
4	No	No	2	

Table 4-14: Survey of production experiments classified by parameters of stimuli and orthography

Although many methodological problems identified in past research have been avoided in recent studies, it seems to me that the presence of minimal pairs, whose inclusion in the test material is standardly thought to be essential for a more controlled, thus comparable, phonetic environment, can exert a confounding influence on the outcome of the experiments. Lisker and Abramson (1967) suggest that an 'enhancement effect' occurs when minimal pairs are involved in the test material. Similarly, Baese and Goldrick (2006) and Baese et al (2007) show that the laryngeal distinction in English obstruents is enhanced in words forming minimal-pair relations involving that contrast. In other words, the existence of minimal-pair neighbours in the lexicon can influence the realisation of the contrast in question.

In most of these experiments, subjects were allowed to go over the list and familiarise themselves with the materials before the beginning of the test. Although such a warm-up might be necessary, why practice lists should be the actual test lists is not at all clear⁵⁸ to me. Interestingly, Snoeren et al (2006) report that incomplete neutralisation is even greater when subjects practise on the test materials at the beginning of the experiment. During practice time, subjects might notice the presence of minimal pairs and might also know that if they do not pronounce the relevant pairs carefully, a lexical ambiguity can arise (unless there are contextual disambiguation clues).

Sentential context has also been found to influence the production of words. More specifically, words that are predictable from context are articulated less carefully.

⁵⁸ Even in situations where stimuli contain rare or unfamiliar items, the experimenter can discuss the meaning of these items with the subjects but should still give them a different list for practice.

For example, the production of these words involves a fair amount of reduction that can undermine their recognisability when excised out of context (Fowler & Housum 1987; Hunnicutt 1985; Lieberman 1963; Meunier et al 2006). Importantly, in studies of neutralisation, the same trend (here, negligible incompleteness) has been reported for pairs produced in contexts providing disambiguating semantics (e.g., Charles-Luce 1993; Port & Crawford 1989).

4.3.1.2 *Experimental Variables*

The discussion above indicates that certain factors have nuisance potential if left untreated. An adequate experimental design must make provision for these factors by experimentally manipulating them as explanatory variables, for instance, or (at least) by controlling for them. In Table 4-15 below, I give a preview of the experimental variables in the current study. These variables are relevant to stimuli and tasks, which will be discussed in turn.

Variable	Type	How is it treated?
Orthography	Independent (to be manipulated)	Two tasks presumably demanding different amounts of attention to orthography
Presence of minimal pairs	Independent (to be manipulated)	Two stimulus lists containing minimal pairs whose members are (1) kept apart and (2) in succession
Sentential context	Independent (to be manipulated)	As part of the task: presence or absence of contextual clues
Prior practice	To be controlled	Not allowed; filler tasks recorded in intervals
Token frequency	To be controlled	Members of each pair are matched for token frequency based on subjective judgments

Table 4-15: Experimental variables in the paradigm

4.3.1.2.1 **Stimuli**

In terms of composition, the stimulus set is the same as described in §4.2.1.2.2. In terms of presentation order⁵⁹, however, there are two stimulus lists, described in (43) and schematised in (44).

⁵⁹ All the five literate speakers in this experiment are given the same presentation order. In other words, the order used in the experiment is not counter-balanced. This is done to minimise inter-speaker variation, which might obscure intra-speaker effects that are due to the experimental factors manipulated in the study. A pilot has shown that these effects are very small. Also, in a small-n study as mine, keeping the sample of speakers as homogenous as possible is of paramount importance. The same is true of the order of presentation of pair members, where words with

- (43)
- i. Pair members apart: minimal pairs are quasi-randomised in such a way that the two members of each pair are not to be found near each other. For example, the list will contain **lah/a/m** and **lah[a]m** but with as many items separating them as possible.
 - ii. Pair members close: minimal pairs are arranged in such a way that the word with $V2_{epenthetic}$ is to follow its $V2_{lexical}$ competitor. For example, the list will contain **lah/a/m** followed immediately by **lah[a]m**.

(44)

i

·
FILLER
FILLER
lah/a/m
FILLER
FILLER
FILLER

·
FILLER
FILLER
FILLER
FILLER

·
FILLER
FILLER
FILLER
lah[a]m
·

ii

·
FILLER
FILLER
lah/a/m
lah[a]m
·
·

4.3.1.2.2 Tasks

In this section, I discuss two more variables from Table 4-15 above. These are orthography and sentential context. The experimental paradigm consists of an elicitation and two reading tasks. In the elicitation task, speakers answer orally

$V2_{lexical}$ invariably precede those with $V2_{epenthetic}$ (see the schematic in the text above). In a larger-n study, it is, of course, preferable to have presentation orders counter-balanced. Importantly, as far as the current study is concerned, the fact that the analyses of 'a' and 'i' do not yield the same results indicates that the non-counter-balanced presentations have not confounded the results in any way.

presented questions. These questions are designed to elicit the target words. In the reading tasks, speakers read out stimulus items presented on a computer screen. The two reading tasks differ with respect to the semantics of the carrier sentence. In the reading-in-context task, each target word appears in a semantically composed sentence, providing disambiguating contextual clues to the meaning of the pair members. In the reading-in-a-frame task, each target word is inserted into the semantically neutral sentence [ga:lat ____ tara] 'she said ____ I think].

The orthographic status of the vowel/zero contrast in BHA needs to be discussed in more detail. BHA is essentially a spoken dialect. The written language for BHA speakers is Standard Arabic (SA). Importing the orthographic tradition of SA to give a written form to the words from BHA does not wholly resolve the vagueness surrounding the contribution of orthography to the phonetics of what is essentially an unwritten dialect like BHA. See the discussion section for more on this.

The Arabic writing system is such that only consonants and long vowels are represented. Short vowels rarely make it into written texts. Only verses of the Qur'an are fully vowelised. Apart from that, written texts appear with short vowels unmarked except sometimes in the case of lexical ambiguities that are not structurally resolved. And even there, short vowels are kept to a minimum in that only the short vowels that are essential to disambiguation are represented as diacritic marks underneath or above the relevant consonants. Other occasions where short vowels are represented include poetry volumes and textbooks taught during early years of school, where pupils are introduced to the writing system in its entirety.

It is only on these occasions that the vowel/zero contrast is represented orthographically. For example, a diacritic symbol beneath the consonant immediately preceding /i/ in the pronunciation, known as 'kasra' in Arabic tradition, represents the lexical short vowel /i/. So, [hɪl] is usually given the Arabic form حل 'state of staying' with no diacritic marking /i/. Very seldom do speakers of Arabic come across the vowelised version حِل with the kasra appearing beneath → [h]. The symbol for lexical short /a/ is [َ], 'fatha', as in حَل [hal] 'solution'. Lexical /u/ is represented by 'dhamma' [ُ], as in حُل [hul] 'find a solution'. Importantly,

there is a symbol for no-vowel (i.e., zero) known as ‘*skuun*’ [ʔ], as in *وَحْل* [wahl] ‘mud’.

However, it seems very unlikely that speakers of an unwritten dialect actually resort to an orthographic representation as complicated as this when they access a vernacular word. There is psycholinguistic evidence that, in Arabic reading, adding short-vowel symbols to written words both disturbs reading processes and delays lexical access (Roman & Pavard 1987). Even the standard only-consonants-and-long-vowels writing system is visually complex. For example, Eviatar et al (2004) suggest that, where Hebrew characters are recognised equally well by both hemispheres, many Arabic characters are indistinguishable by the right hemisphere. Ibrahim et al (2002) conclude that Arabic orthography with multiple symbols for any single consonant⁶⁰ and with many confusable symbols used for different consonants, not to forget the famous disregard for short vowels, actually slows the processing of the characters and hence the recognition of written words. This also finds support in findings by Roman and Pavard (1987: 439), who report that Arabs need to look at Arabic words printed in Arabic characters 1.5 times longer than French need to look at French words printed in Roman characters. The researchers conclude that Arab readers need to extract more information from the printed text.

The cognitive difficulty involved in the processing of Arabic orthography points towards what must be a very marginal place for orthography in neutralisation. It is hoped that elicitation and reading tasks should place different demands on speakers to pay attention to orthography. It is by no means obvious how the putative effect of orthography can be eliminated in an elicitation task with literate subjects, who know how the target words are spelled in Standard Arabic, and might access an orthographic representation when retrieving the word in response to interview questions. Nonetheless, as BHA is not written, an elicitation task should draw less attention to orthography than would a reading task where the stimuli are presented in writing.

⁶⁰ Eviatar et al (2004: 175) observe that out of the twenty-eight characters used to represent consonants in Arabic writing, twenty-two “have four shapes each” according to where they occur in the word and/or phrase.

4.3.1.3 *The Paradigm*

The experimental paradigm in this study is a two (V₂-UR-Status) by two (stimulus lists) by three (tasks) factorial design. Epenthetic and lexical data alike are collected in six experimental conditions manipulating stimulus materials and tasks, as illustrated in Table 4-16.

	Condition	Stimulus list	Sentential context	Task
Block I	1	Members apart	Composed sentence	Elicitation
	2	Members apart	Composed sentence	Reading
	3	Members apart	Frame sentence	Reading
Block II	4	Members in succession	Composed sentence	Elicitation
	5	Members in succession	Composed sentence	Reading
	6	Members in succession	Frame sentence	Reading

Table 4-16: Experimental conditions in blocks based on stimulus list

One of the variables in the paradigm is the presentation order of pair members in the stimuli. If the stimuli are going to be recorded six times by each literate speaker, there is a genuine need to control for any learning effects. But with only five literate speakers available for recording, traditional methods of having different speakers do different conditions in different orders can foster a carry-over effect for those speakers who do condition 6 in Table 4-16 before they do condition 1, which is the least contrastive. Condition 6 is highly contrast-promoting in terms of stimulus presentation. In terms of orthography, however, it is highly neutralisation-promoting.

The strategy I adopted in the production experiment reported here was not to provide an orthography at the outset of the experiment. Rather, I decided to start with the least ostentatious condition, which involves eliciting an unwritten list of words that contains minimal pairs whose members appeared as far apart as possible. As to what conditions to follow, there appear to be two options—one being to move to the next stimulus list after all combinations of the previous list and tasks have been recorded (i.e., to have stimulus-oriented blocks of conditions), the other being to continue with the same task while varying stimulus lists until all the relevant combinations have been recorded, then move to the next task and try both stimulus lists, etc. (i.e., to have task-oriented blocks). A third possibility was to give up on having a within-speaker design and recruit five groups of speakers where each group completes only one block of conditions—two groups do the two

stimulus blocks (i.e., members-apart conditions and members-close conditions); three groups do the three task blocks (i.e., elicitation conditions, reading-in-context conditions, and reading-in-a-frame conditions). Here, elicitation can be the opening condition for the first two groups of speakers; so can members-apart-list for the other three groups. What we would sacrifice instead is the one advantage that a within-speaker design guarantees—having each speaker as her own control.

Practical limitations on this thesis dictated having blocks done by the same speakers. But here the decision I made was to have stimulus-oriented blocks. Accordingly, across all conditions, members-apart stimulus lists were recorded first, while within each block, the elicitation task was always the first. The decision to give priority to stimulus lists over tasks was mainly based on the insight emerging from the survey in Table 4-14 that the presence of minimal pairs has reasonable potential to influence the phonetics of neutralisation. In contrast, based on findings of psycholinguistic studies, I suggested in §4.3.1.2.2 above that orthography stands a slight chance of confounding production data from an unwritten dialect like BHA.

Furthermore, the following methodological strategy was applied during the execution phase of the production experiment. Recording sessions were scheduled to take place on different days, with a few days' interval separating any consecutive conditions. During those intervals, speakers participated in what can be described as filler recordings, such as reading out wordlists, telling stories, and answering questions in an interview. All of these activities have different materials from what is used for the current experiment.

4.3.1.3.1 Method

4.3.1.3.1.1 Speakers

The data come from the five literate speakers whose details are given in §4.2.1.2.1 above.

4.3.1.3.1.2 Materials

The test materials are the same as described in §4.2.1.2.2 above. See also §4.3.1.2.1. A total of 2520 tokens (14 pairs (14 x 2) x 3 repetitions x 2 stimulus

lists x 3 tasks x 5 speakers) were acoustically analysed using Praat (Boersma & Weenink 2008).

4.3.1.3.1.3 Procedures

The data in this experimental paradigm were acquired following the procedures detailed in §4.2.1.2.3. As explained before, the elicitation tasks proceed as an oral interview conducted for each speaker individually. During those interviews, speakers responded to the orally presented questions, saying the target word in the frame [__tara]. In the reading-in-context tasks, speakers read out each sentence, as it appeared on a computer screen. The sentential context provides disambiguating clues to the meaning of pair members and is generally of the form [*phrase*__tara]. This semantically disambiguating phrase is necessarily different for different words. Finally, in reading-in-a-frame conditions, speakers read out the target words in the frame [ga:lat__tara] 'she said__ I think'. This frame sentence appeared only once at the beginning of the exercise. Each target word appeared on the computer screen accompanied by a word in parentheses indicating the morpho-syntactic class of the target item. Speakers were instructed to look at both the target item and its accompanying descriptor (noun or verb), which appeared in red, before they say the target item in the frame. Speakers were not allowed to practise on the test words. They had a different practice list to familiarise themselves with the procedures.

As to statistical procedures, it serves the goal of this chapter to maximise the comparability of the statistical treatment of the data analysed here with those reported in the literature. This is because the chapter attempts to shed light on the genuineness argument as it has evolved and been defended in the literature.

4.3.1.4 Results

The general picture emerging for the analysed data is that differences between epenthetic and lexical vowels, when found, are exceedingly small. The magnitude of differences and variability within the data are not the same across the different conditions. Appendix G gives the mean and SD values of epenthetic and lexical vowels along the five acoustic measures investigated here. These values are summarised by conditions, stimulus lists, and tasks.

Overall, participants produced larger differences between epenthetic and lexical vowels in stimulus conditions with pair-members apart than in those with pair-members close. This is the case for most measures for 'a' and 'i'. Exceptions include i-duration and i-F2. In contrast, the contrast-promoting conditions where speakers produce members of minimal pairs in succession generally show the smallest differences.

The smallest variation appears in the elicitation tasks. Specifically, condition 1 is by far the least variable. The largest variation is found in the conditions where the target words are read out in disambiguating sentences—conditions 2 and 5. Of these, condition 5 is by far the most variable. All of these observations apply to most parameters for both 'a' and 'i'. Interestingly, condition 6, where pair members which appear in succession are read out in a semantically neutral frame, shows constrained variation in F0 and duration for both 'a' and 'i'. Mean and SD data are graphed in Figure 4-17 through Figure 4-26.

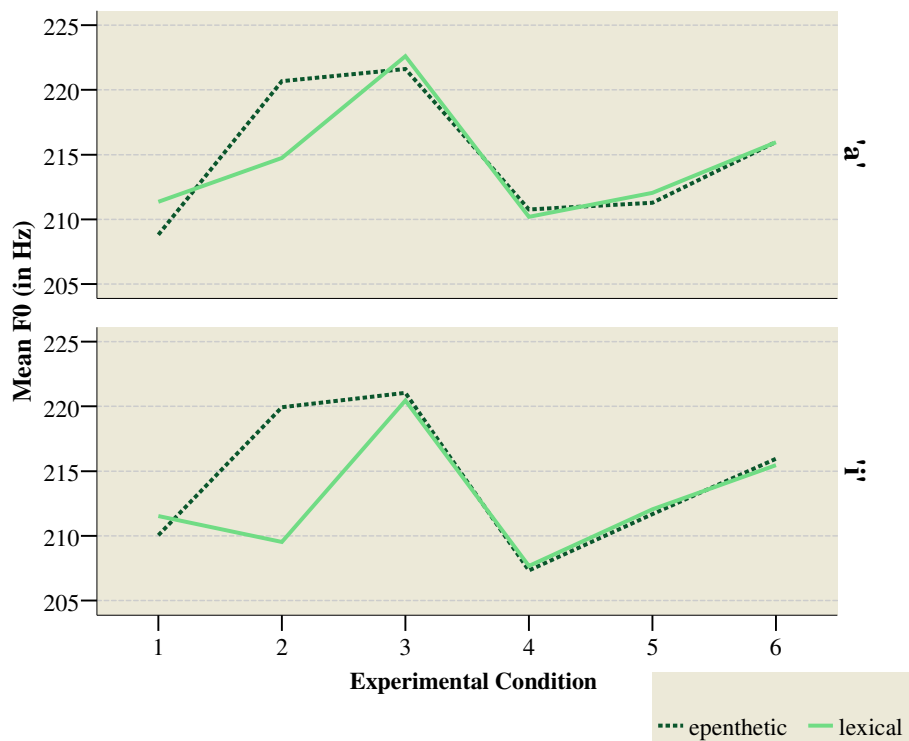


Figure 4-17: Mean F0 by experimental condition for 'a' and 'i'

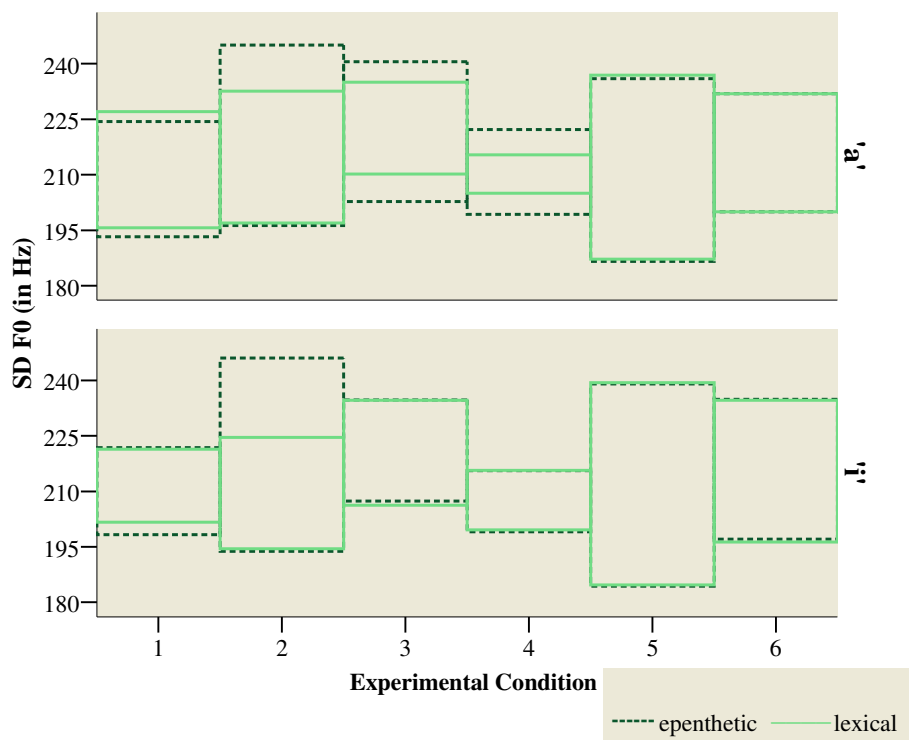


Figure 4-18: F0 SD by experimental condition for 'a' and 'i'

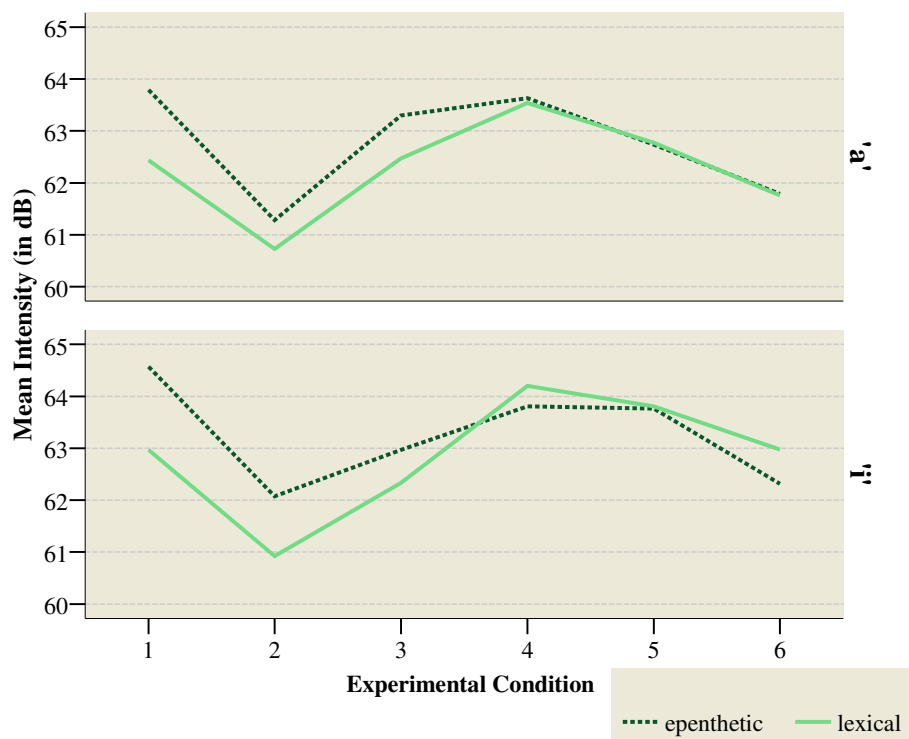


Figure 4-19: Mean intensity by experimental condition for 'a' and 'i'

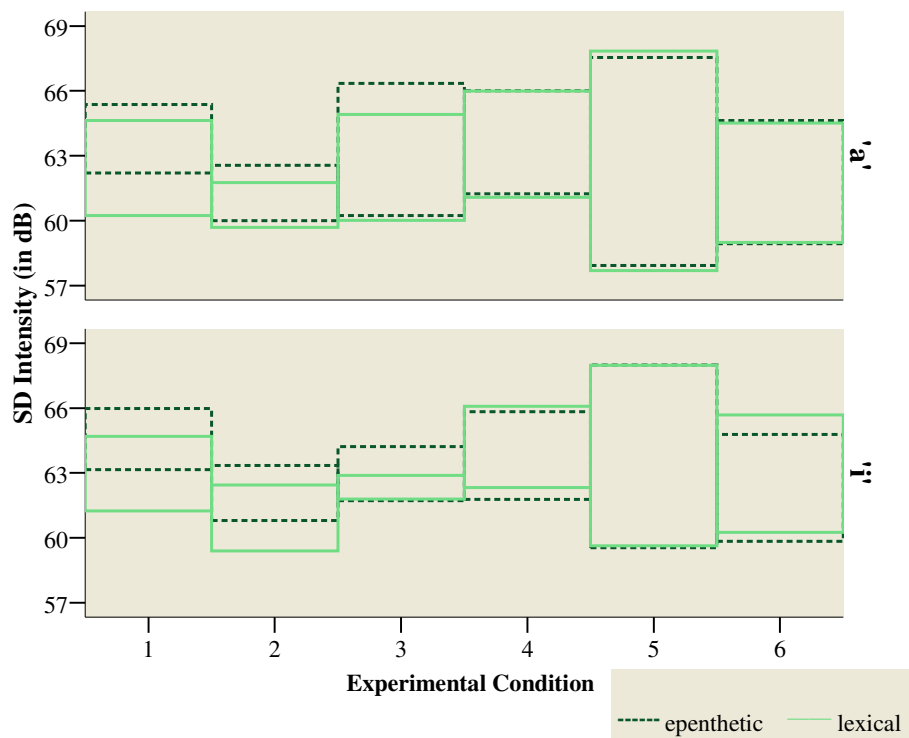


Figure 4-20: Intensity SD by experimental condition for 'a' and 'i'

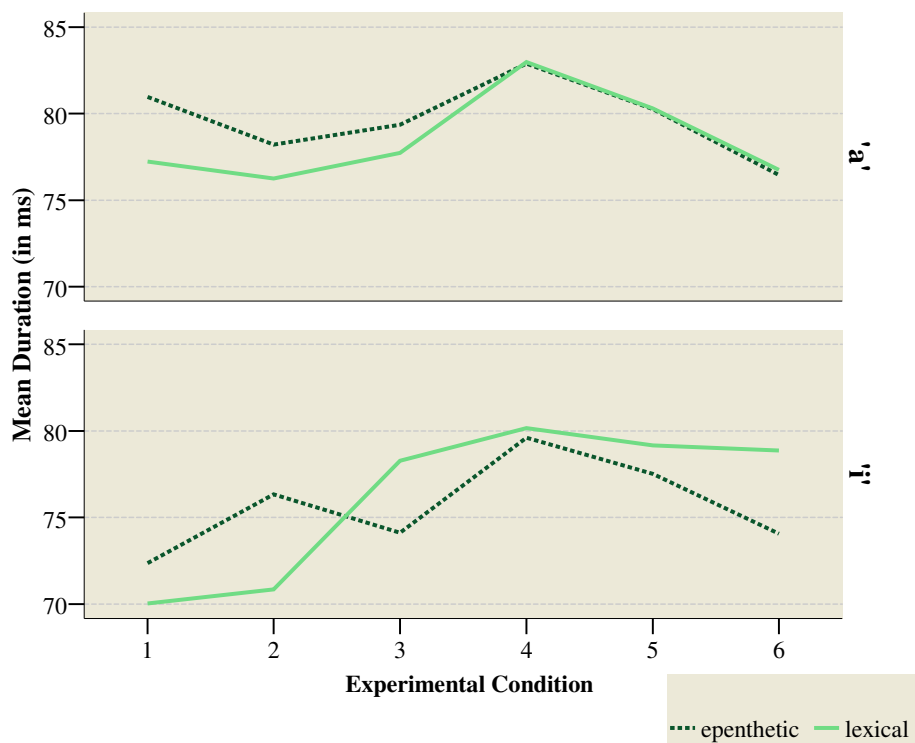


Figure 4-21: Mean duration by experimental conditions for 'a' and 'i'

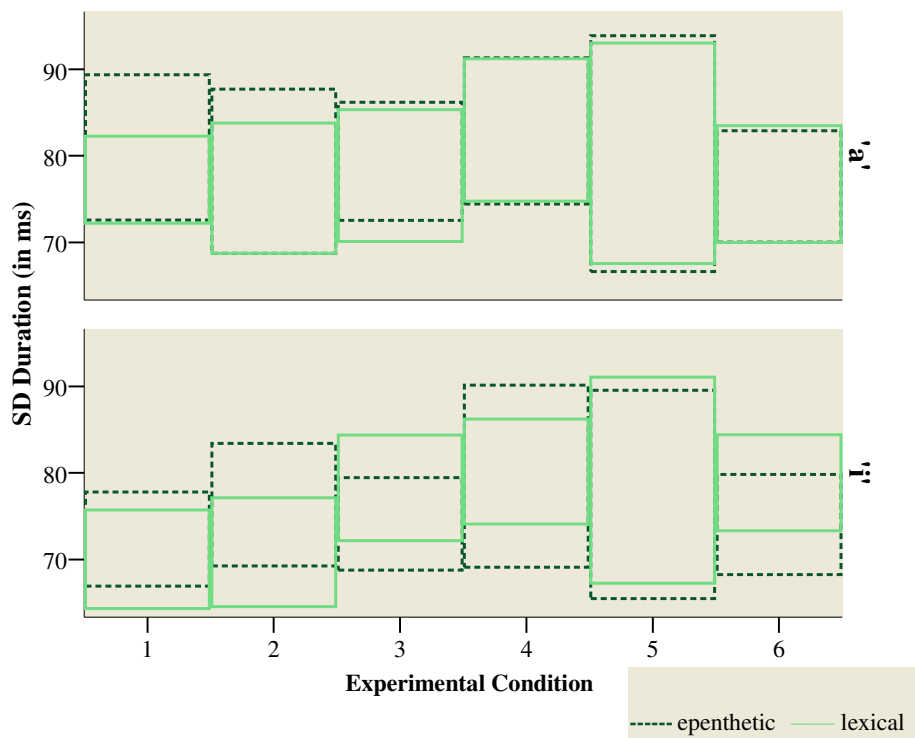


Figure 4-22: Duration SD by experimental condition for 'a' and 'i'

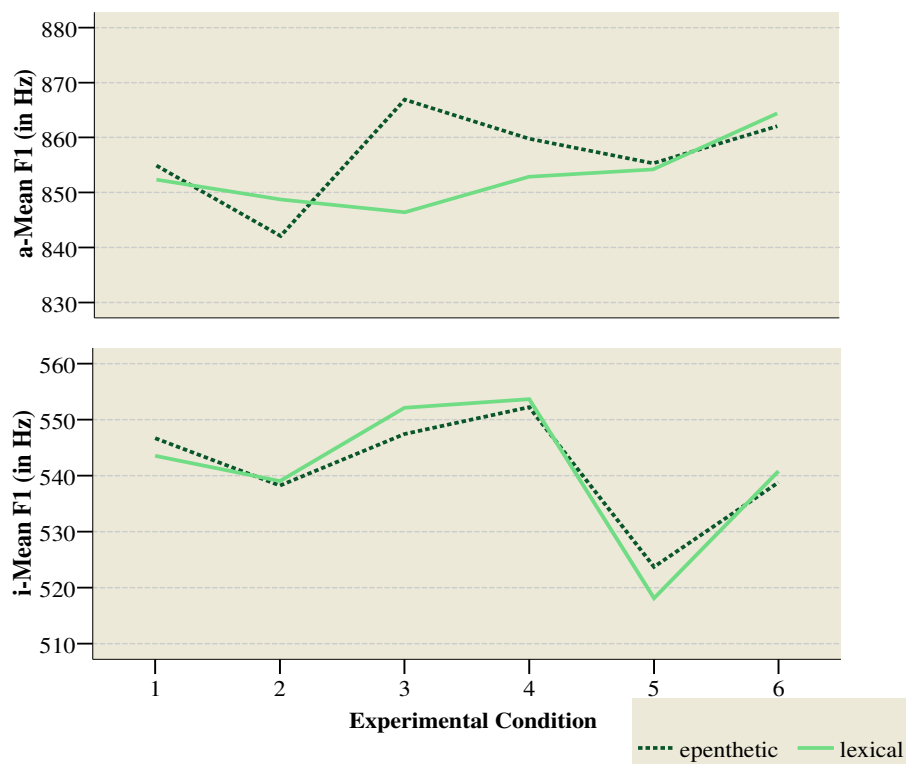


Figure 4-23: Mean F1 by experimental condition for 'a' and 'i'

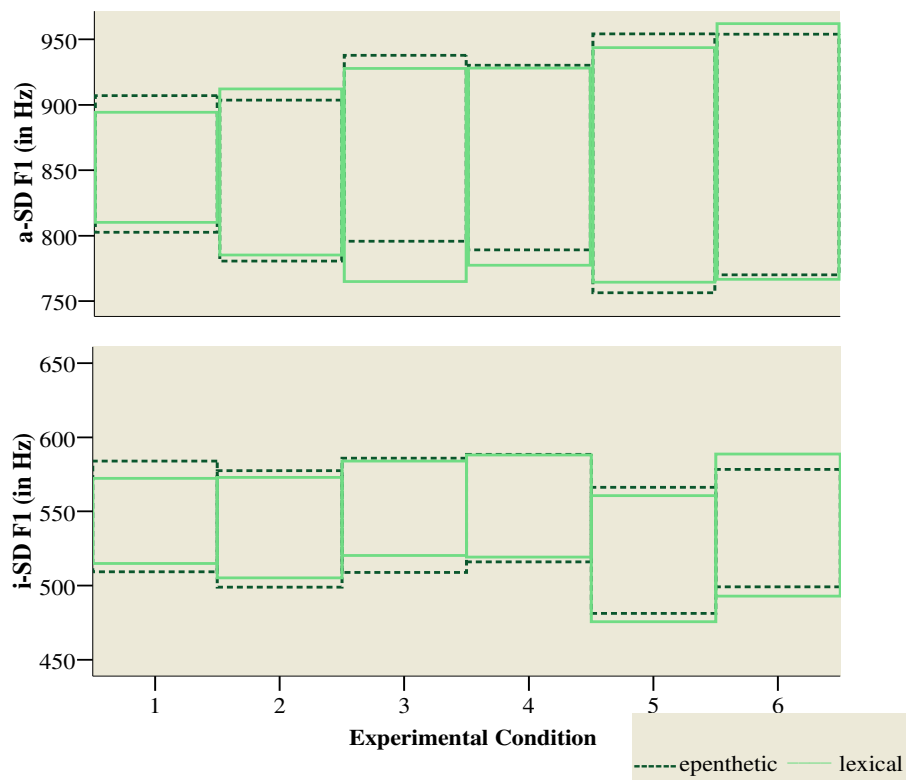


Figure 4-24: F1 SD by experimental condition for 'a' and 'i'

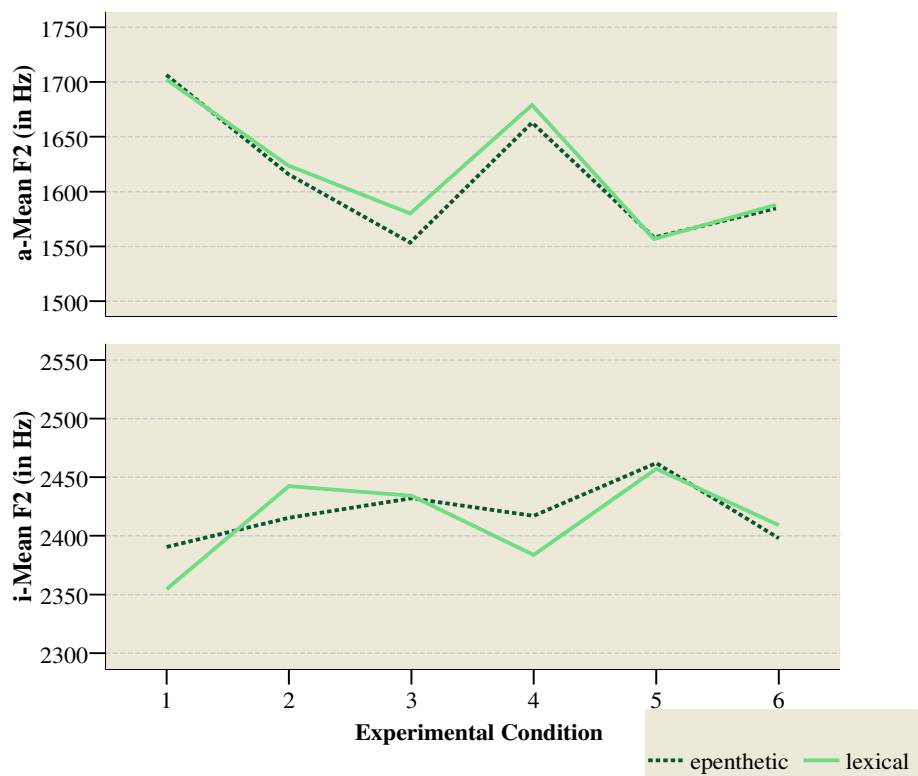


Figure 4-25: Mean F2 by experimental condition for 'a' and 'i'

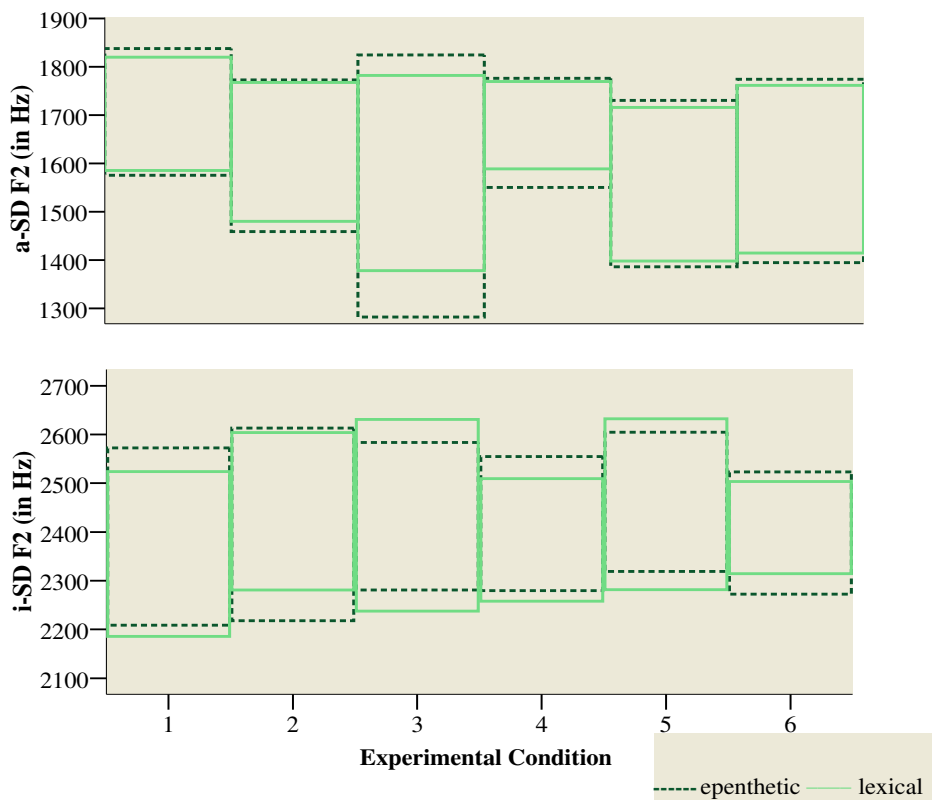


Figure 4-26: F2 SD by experimental condition for 'a' and 'i'

Mean differences as described above are more clearly captured in Figure 4-27 through Figure 4-31, which graph mean paired differences (\bar{X}_{PD}) and SD_{PD} . These summary-statistics values appear in Appendix H.

In terms of variation, however, these figures present a slightly different picture of the data than the absolute SD values above. For example, while most acoustic measures for both 'a' and 'i' data display the largest variation in the reading-in-context tasks, the magnitude and direction of the paired differences along most of these parameters show very little variation in these tasks. In other words, production varies a lot from the group mean value, but not from the group mean paired difference. Variation in the first case is segmentally based. Being placed in necessarily different segmental contexts, vowels in the target words are predictably produced with a lot of variation. But speakers apparently produced more or the less the same difference (rather, non-difference) between epenthetic and lexical vowels in these conditions.

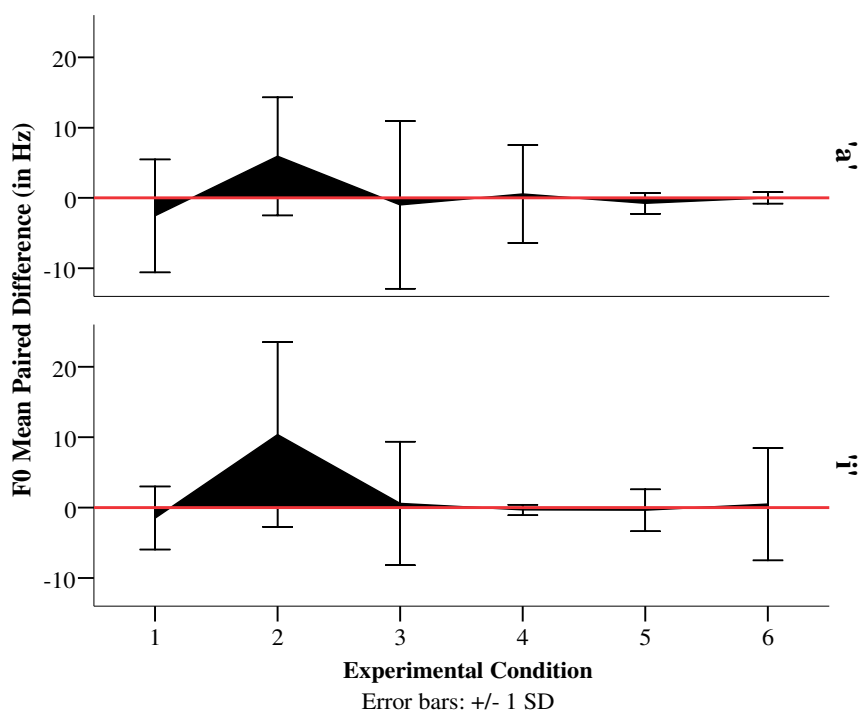


Figure 4-27: F0 mean paired differences \bar{X}_{PD} and SD_{PD} values (epenthetic - lexical) according to experimental conditions for 'a' and 'i'

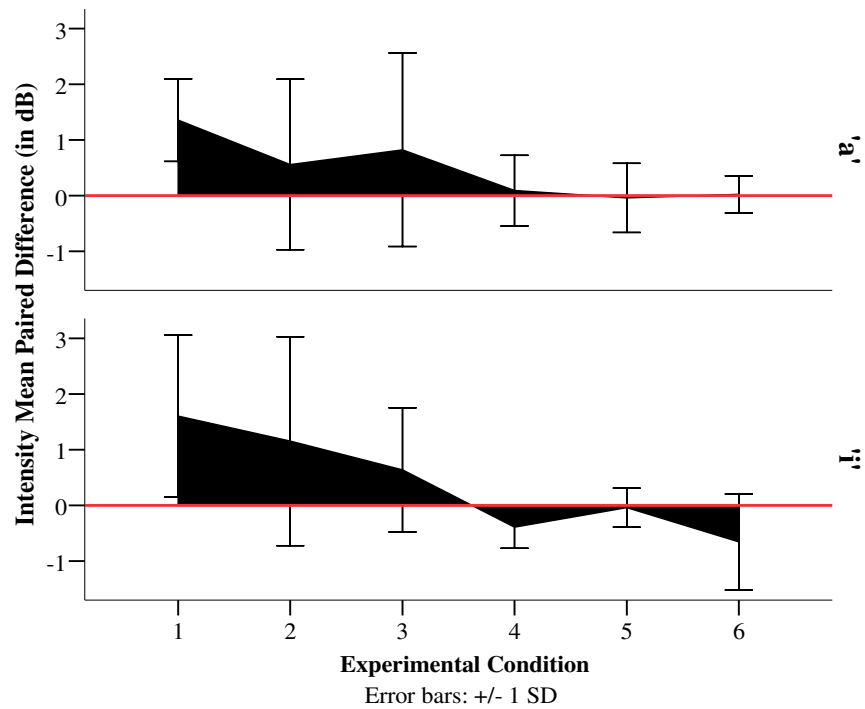


Figure 4-28: Intensity mean paired differences \bar{X}_{PD} and SD_{PD} values (epenthetic – lexical) according to experimental conditions for 'a' and 'i'

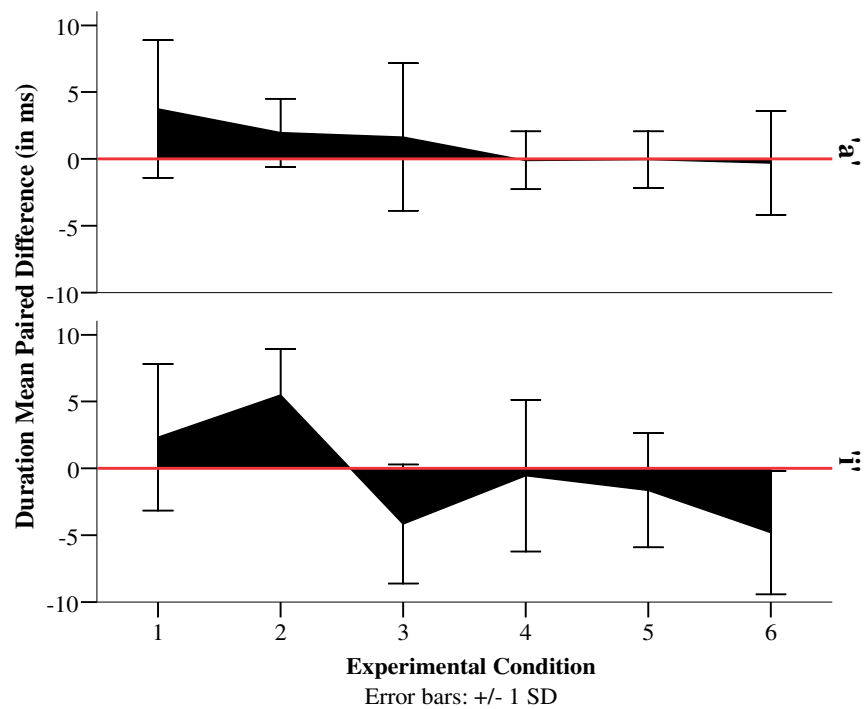


Figure 4-29: Duration mean paired differences \bar{X}_{PD} and SD_{PD} values (epenthetic – lexical) according to experimental conditions for 'a' and 'i'

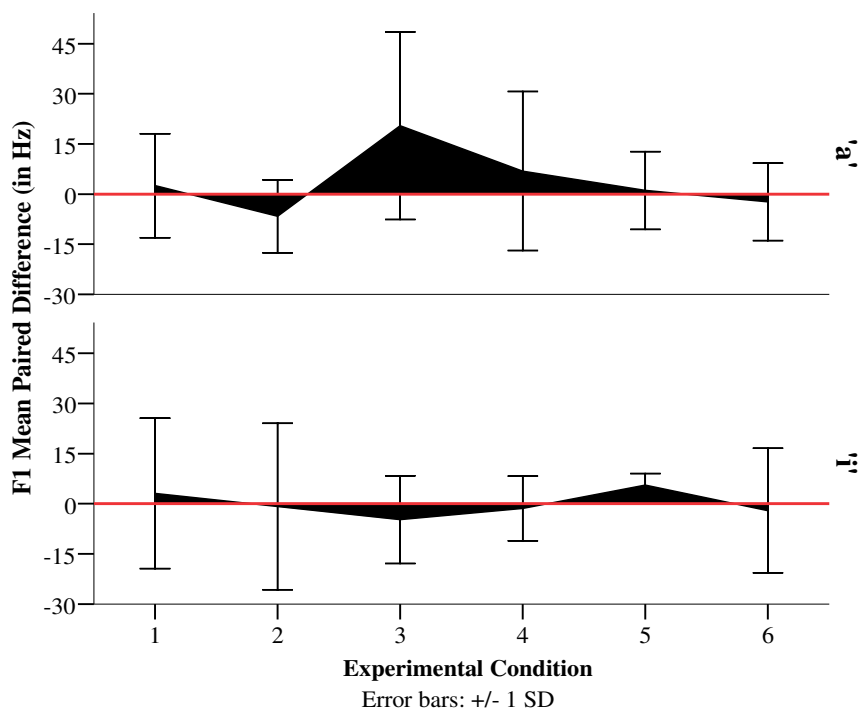


Figure 4-30: F1 mean paired differences \bar{X}_{PD} and SD_{PD} values (epenthetic – lexical) according to experimental conditions for 'a' and 'i'

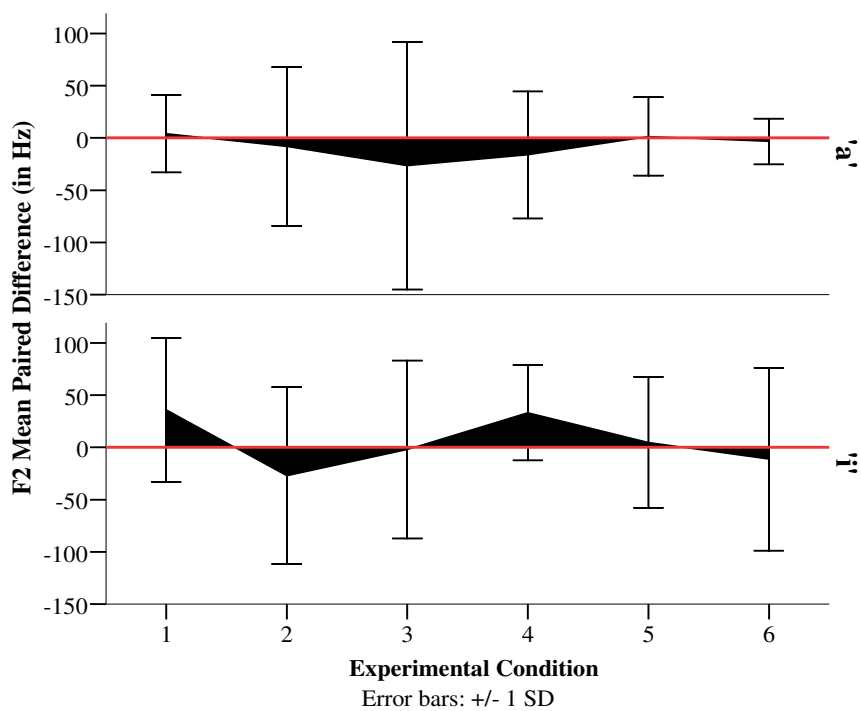


Figure 4-31: F2 mean paired differences \bar{X}_{PD} and SD_{PD} values (epenthetic – lexical) according to experimental conditions for 'a' and 'i'

Looking at the box-plots of the data in Figure 4-32 through Figure 4-36, we may notice different effects of stimulus and task manipulations on the acoustic measures. These effects are summarised in Table 4-17 below.

Acoustic Measure	Stimuli	Task
F0	Smaller range in members-close list data	Elicitation the least variable; reading in context the most variable
Intensity	Less variability in members-apart list	Elicitation the least variable; Condition 5 the most variable
Duration	Less variability in members-apart list	Reading in context the most variable; reading in a frame the least variable for 'a'; elicitation the least variable for 'i'
F1	No obvious pattern	Lower i-F1 in reading in context
F2	No obvious pattern	Elicitation the least variable for 'a'; no obvious pattern for 'i'

Table 4-17: Stimulus and task effects on acoustic measures for 'a' and 'i' based on the data summarised in the box-plot graphs below

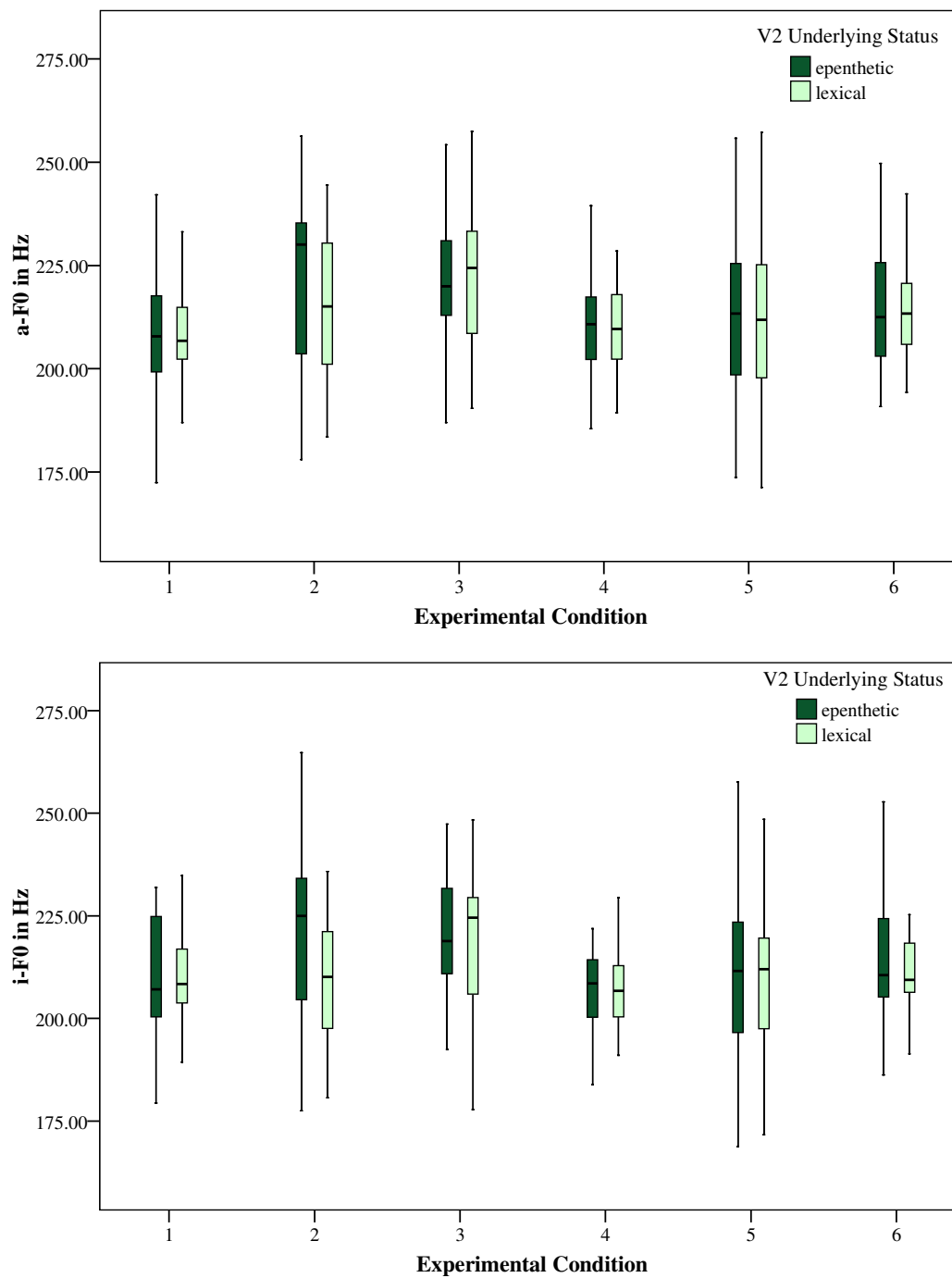


Figure 4-32: Box-plots for a-F0 and i-F0 according to V₂ Underlying Status by experimental condition; the graphs show the median, upper and lower quartiles, and range of 'a' and 'i' data.

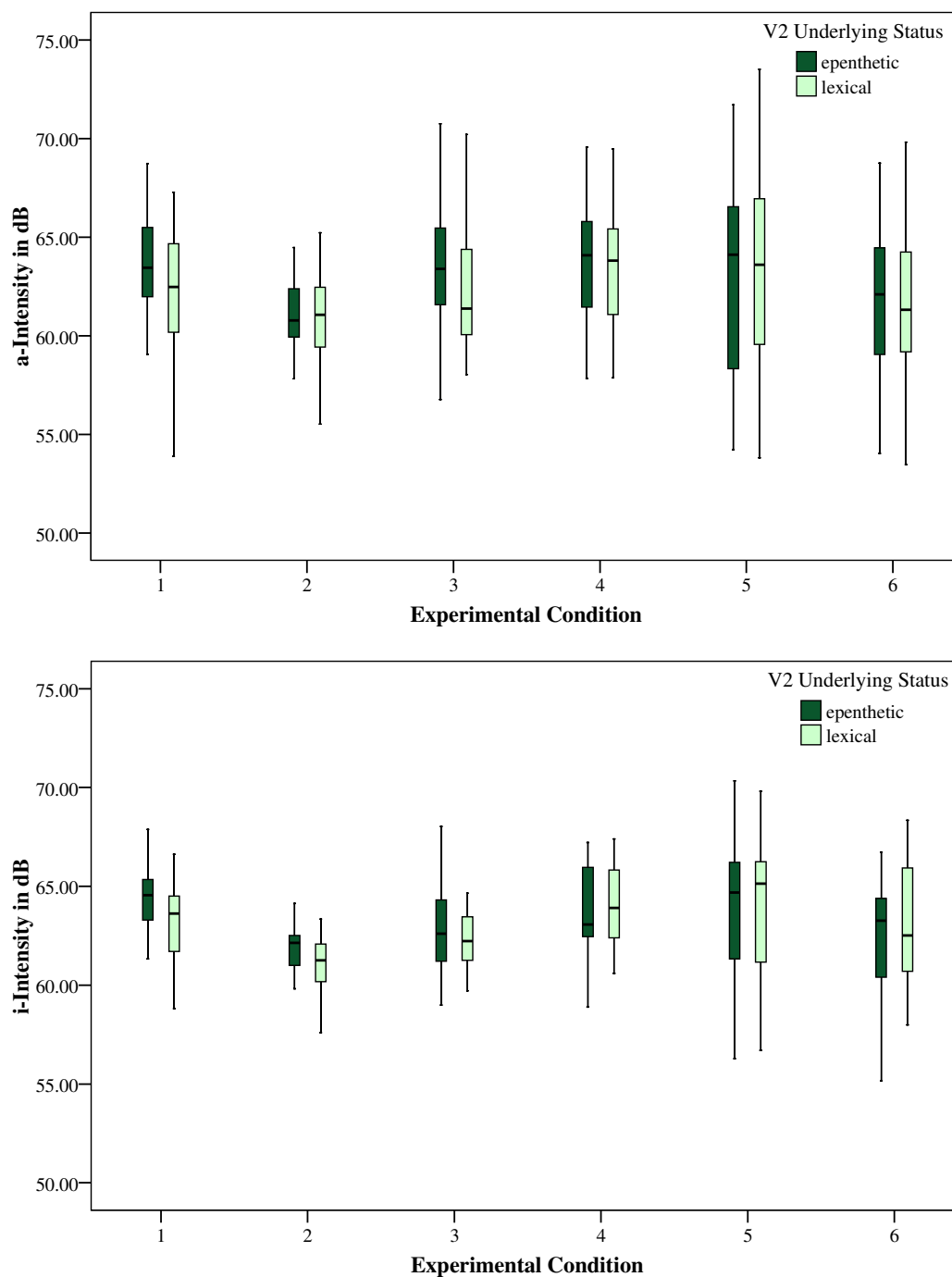


Figure 4-33: Box-plots for *a*-intensity and *i*-intensity according to V_2 Underlying Status by experimental condition; the graphs show the median, upper and lower quartiles, and range of 'a' and 'i' data.

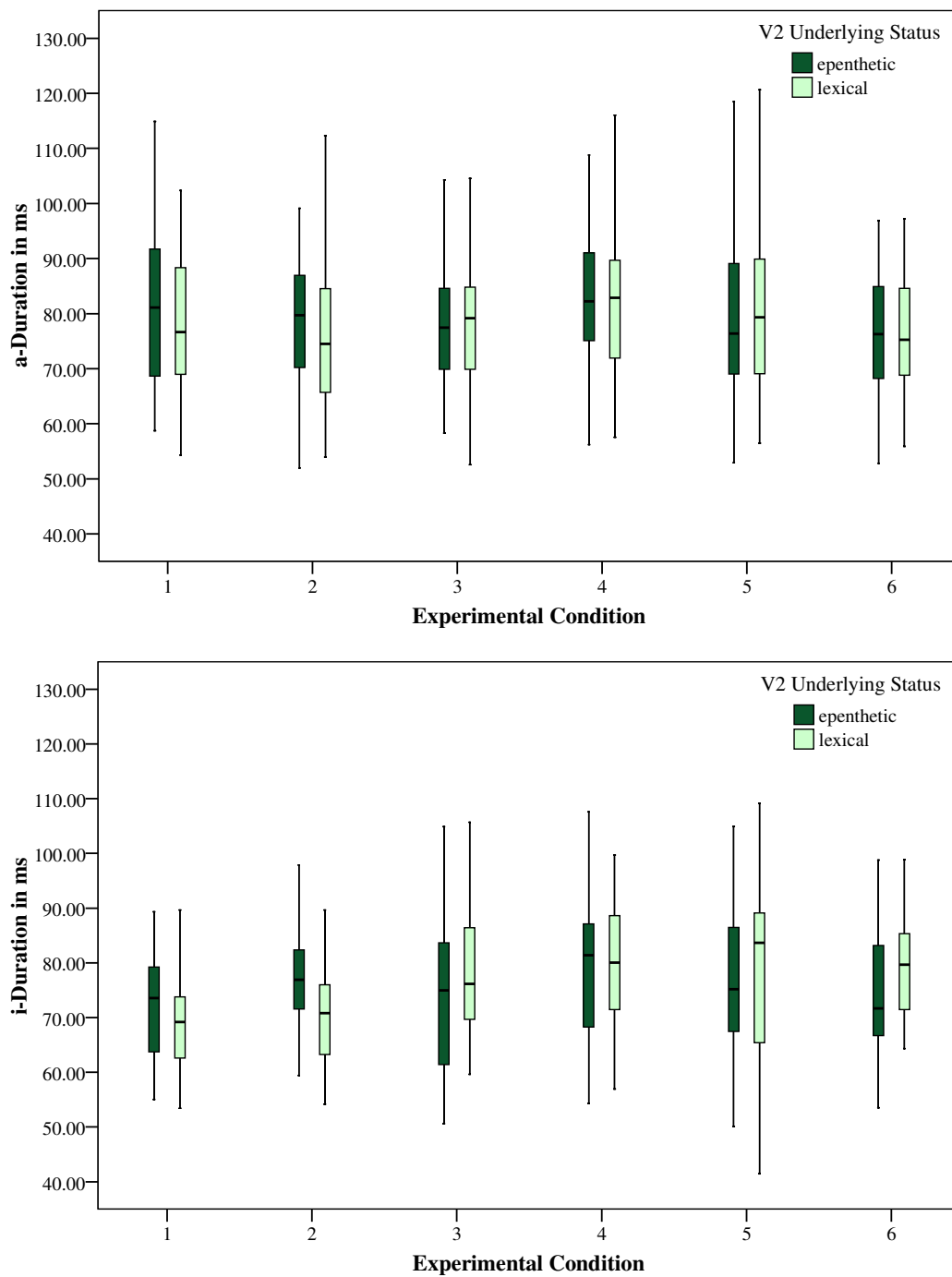


Figure 4-34: Box-plots for a-duration and i-duration according to V_2 Underlying Status by experimental condition; the graphs show the median, upper and lower quartiles, and range of 'a' and 'i' data.

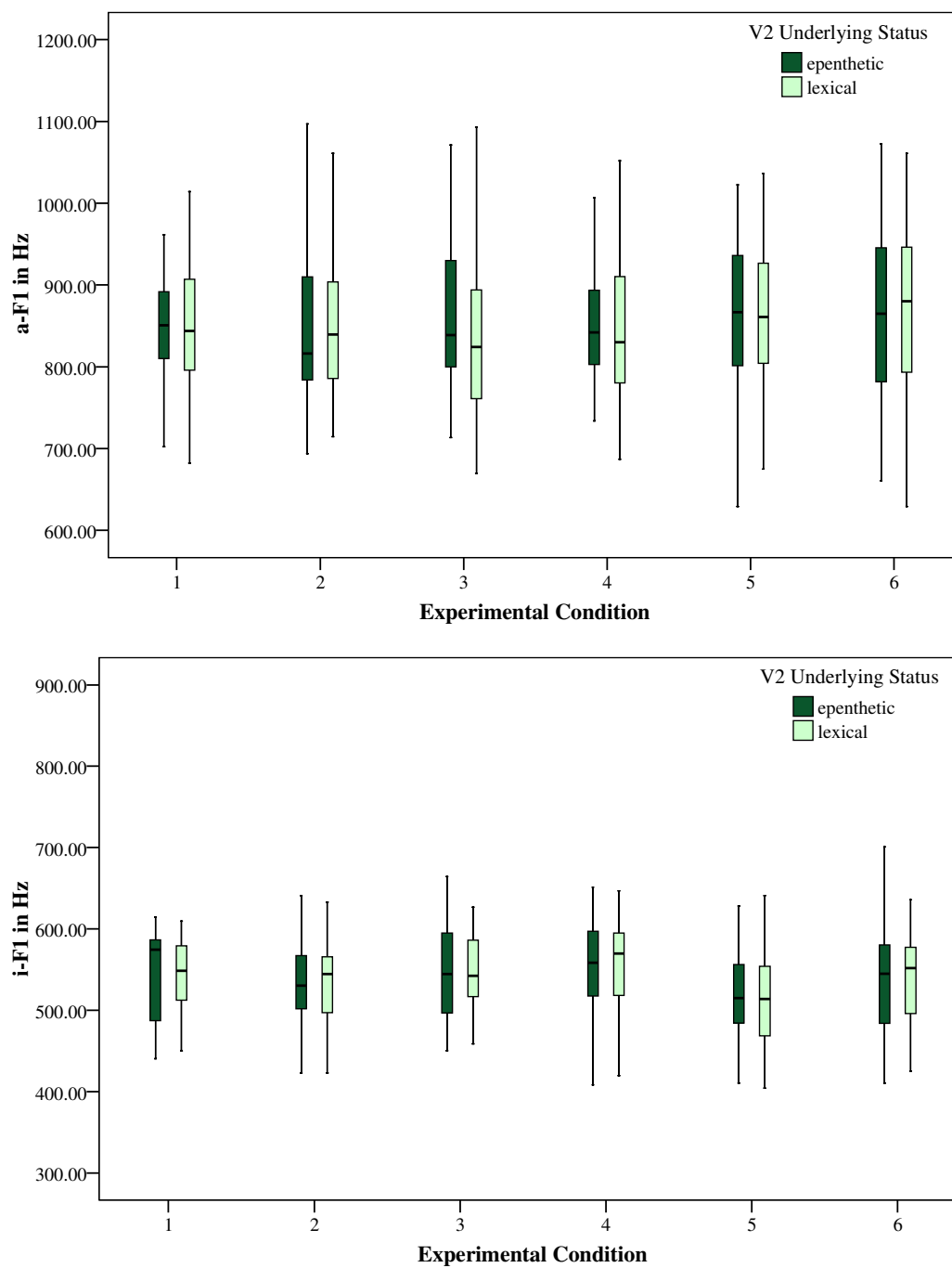


Figure 4-35: Box-plots for *a-F1* and *i-F1* according to V_2 Underlying Status by experimental condition; the graphs show the median, upper and lower quartiles, and range of 'a' and 'i' data.

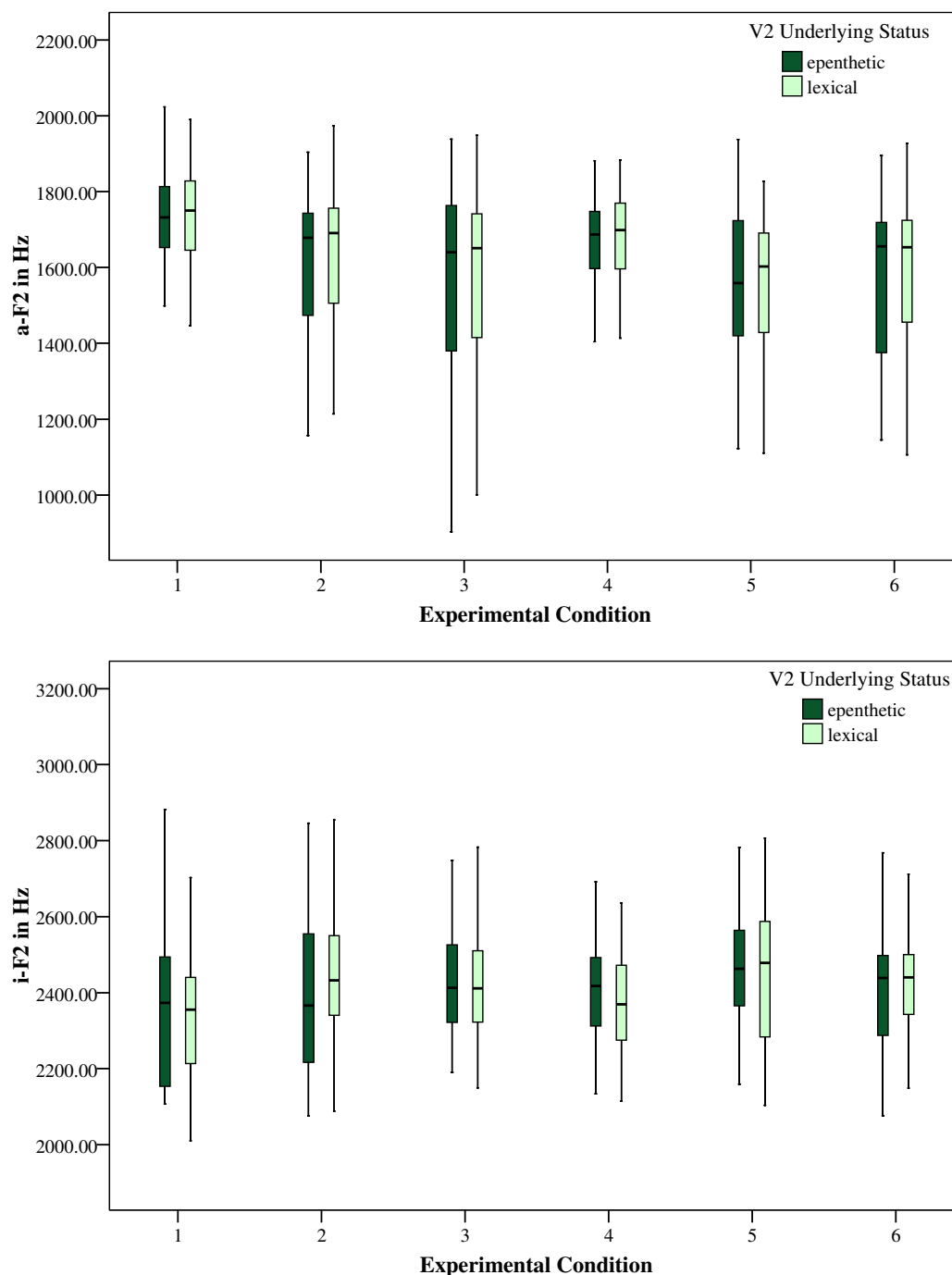


Figure 4-36: Box-plots for *a*-F2 and *i*-F2 according to *V*₂ Underlying Status by experimental condition; the graphs show the median, upper and lower quartiles, and range of 'a' and 'i' data.

A series of repeated-measures Anovas were run separately for each of the five acoustic parameters. Anovas with *V*₂-UR-Status (epenthetic vs lexical), stimulus list (members-apart vs members-close), and task (elicitation vs reading-in-context vs

reading-in-a-frame) as factors found no statistically significant main effects nor interactions along any parameters for 'a'.

Similar Anovas, however, reveal that for 'i' there is a statistically significant main effect of task on i-F1 ($F(2,8) = 8.86$, $p = .009$) with estimated marginal means as follows: elicitation = 549Hz (SE = 14.6); reading-in-context = 529.7Hz (SE = 16.2); reading-in-a-frame = 544.7Hz (SE = 14.7). Post hoc comparison tests with Bonferroni adjustment reveal that the i-F1 in elicitation tasks is statistically significantly different from i-F1 in the reading-in-context tasks.

The interaction between V_2 -UR-Status and task is statistically significant for i-duration ($F(2,8) = 13.58$, $p = .003$) with estimated marginal means as follows: elicitation = [i]: 76ms (SE = 3.36), /i/: 75ms (SE = 2.45); reading-in-context = [i]: 77ms (SE = 3.95), /i/: 75ms (SE = 3.6); reading-in-a-frame = [i]: 74ms (SE = 2.2), /i/: 78.5ms (SE = 2). A non-parametric Wilcoxon signed ranks test reveals that the durational difference between epenthetic [i] and lexical /i/ is statistically significant in condition 6 ($Z = -2.02$; $p = .04$).

Finally, i-intensity data display a statistically significant interaction involving V_2 -UR-Status and stimulus list ($F(1,4) = 8.2$, $p = .04$) with estimated marginal means as follows: members-apart list = [i]: 63.2dB (SE = .3), /i/: 62 (SE = .35); members-close list = [i]: 63.2dB (SE = 1.2), /i/: 63.6 (SE = 1.2). A non-parametric Wilcoxon signed ranks test reveals that the intensity difference between epenthetic [i] and lexical /i/ is statistically significant in condition 1 ($Z = -2.02$; $p = .04$). See Appendix I and Appendix J for graphs displaying main effects and interactions among the factors for both 'a' and 'i'.

4.3.1.5 Discussion

I discuss the effect of experimental manipulations on the phonetics of neutralisation from the perspective of task and stimulus main effects and their interactions with V_2 -UR-Status.

I discuss the effect of task manipulations with reference to sentential context, orthography and delivery. The experimental paradigm in this study includes an elicitation and two reading tasks. In the elicitation task, each orally-presented question, which provides the disambiguating context, is answered by a two-word clause. The carrier in reading tasks, a composed sentence or a frame, is a three-

word clause. That is, the elicitation task is different from the two reading tasks not only in terms of clause length but also in the fact that the stimulus materials are not presented in an orthographic medium. Given that no data-transformation has been carried out to make up for clause length, any differences observed between elicitation and reading cannot be conclusively attributed to orthography; these can be just as well time-related (but see below).

As shown above, the only statistically significant main effect involves an i-F1 difference between elicitation and reading-in-context. Note that both of these conditions provide disambiguating contextual clues as part of the elicitation question and the carrier sentence, respectively. i-F1 in reading-in-context tasks is the lowest among all three tasks. Since it is only in this task that the carrier sentence varies in segmental, syntactic, and semantic properties, the difference is most probably segmentally conditioned. This finds support in the fact that i-F1 mean values are very close in the other two tasks and considerably higher than i-F1 mean value in the reading-in-context task. Thus, this F1 difference cannot be due to durational differences.

Failure of the other acoustic measures to display statistically significant differences in response to task manipulation indicates that manipulating task in this way has an extremely limited impact on the production of the target vowels in the test words. For example, the three levels of the independent variable of task might be too close together for the subjects to produce a discernible effect (see e.g., Snodgrass et al 1985 for illustrations). More specifically, there might be no, or very negligible, durational differences between a two-word clause and a three-word clause in BHA phonetics. If true, this would effectively remove the confounding potential of the different clause-lengths that have been used. The orthography-based distinction holding between the elicitation task and the two reading tasks co-varies with a clause-length difference. Since task manipulations have largely failed to produce an effect, the obvious conclusion for both orthographic and clause-length differences, as employed in this study, is that neither matters too much for the acoustics of BHA vowels. It is possible that a different manipulation using a different range or set of task levels may produce an effect. I leave this for future research.

It is also important to remember that my dismissal of task manipulation in this experiment as unsuccessful largely rests on the outcome of statistical significance procedures, which I take up further in the next section. Interestingly, the descriptive statistics point to a delivery-based difference. Specifically, in the elicitation task, vowels tended to be produced with a lower F0 and more intensity than in the reading tasks. At the same time, 'a' has a higher F2, while 'i' has a lower F2 in the elicitation task than in the reading tasks. This spectral difference suggests that 'a' is less back while 'i' is less front in the elicitation task than in the reading tasks. A lower F0 and greater intensity often correlate with more openness (e.g., Fischer-Jørgensen 1990; Lehiste & Peterson 1959; Whalen & Levitt 1995; Yi Xu, p.c. 2010). Taken together, these trends characterise a centralised vowel production, at least as far as 'i' is concerned. However, there are at least two reasons to doubt the validity of claiming that the vowels are more centralised in a two-word elicitation task than in three-word reading tasks.

Firstly, as is well-known, a form of formant centralisation, sometimes described as undershoot, correlates with durational reduction (e.g., Lindblom 1963; Lindblom 1990; Moon & Lindblom 1994). In the data at hand, there is no durational reduction, but there appears to be a kind of formant centralisation. Finding centralisation can sometimes simply be a side-effect of calculating group mean values, where inter-speaker variations pull and push the grand mean towards the centre of the data (cf. van Bergem 1995). Although I do not look into individual data for reasons of space and time, I have, nonetheless, investigated this possibility for the F2 data and found that the pattern of centralisation in the group mean generally occurs for individuals' mean values. This implies that the centralisation here is not an artefact of average calculation over speakers. Note, however, that the within-speaker centralisation effect can still be an artefact of average calculation over items. I leave this for a future follow-up study. In chapter six, I propose a model of lexical representation and phonetic processing that offers a more principled way of treating speaker- and item-related variations.

An important question that suggests itself here concerns the prevalence of durational fluctuations in a dialect where vowel length is contrastive (cf. Flemming 1997; Keating 1985). Acoustic studies on Arabic dialects conclude that the duration of short vowels usually resists environmental effects. For example, Mitleb

(1984), and Haj Yusef (1992) report that vowel duration in Arabic does not show an effect of the voicing of the preceding consonants, but see Alghamdi (2004) for a different outcome. Haj Yusef (1992) shows that vowel duration in Arabic does not fluctuate in response to changes involving the surrounding consonantal environment, such as pharyngealisation, gemination, and other place- and manner-of-articulation effects. If this is true of BHA, then the lack of durational reduction does not necessarily invalidate the centralisation claim.

Secondly, the lack of any F1 differences seems to call into question any appeal to the correlation between articulatory openness and lowered pitch and increased intensity that previous research on vowel acoustics has documented. There are at least two explanations for this lack of F1 variation. Firstly, this immunity of F1 may be related to suggestions in the phonological literature that Arabic only has a height contrast in short vowels. That is, /i/ and /a/ are phonemes, while [u] is not (Herzallah 1990; McCarthy 1994; but see Al-Mozainy 1981). Lindholm (1986: 39) notes the “primacy of height (sonority or F₁) over front-back (chromaticity or F₂) distinctions” in vowel systems. In other words, in Arabic the backness dimension is less constrained than the contrastive height dimension. Interestingly, Tsukada (2009) found no effect of vowel length (i.e., long vs short vowels) on the realisation of F1 in Arabic. This is in contrast to Japanese and Thai, both of which employ a length contrast just like Arabic. Furthermore, this agrees with findings reported in de Jong and Zawaydeh (1999) that F1 differences between stressed and unstressed vowels are very small ranging from 10Hz to 40Hz. Secondly, it is argued in the literature that the production of /i/ and /a/ is subject to a saturation effect causing muscular tensions beyond a certain point to have no articulatory or acoustic consequences (e.g., Perkell 1996). The small variability in i-F1 and a-F1 is probably a manifestation of this effect.

In any case, the magnitude of the raw mean differences is very small, anyway. The ratio holding between the mean values of any two tasks along all the acoustic measures for both ‘a’ and ‘i’ broken down by V₂-UR-Status approximates ≈1 (see Appendix G and Appendix H). Clearly, more research is needed to give a complete account of this.

Next I discuss the interaction of task with V₂-UR-Status, which provides a more direct window into the phonetics of vowel/zero neutralisation in BHA. Here, i-

duration provides the only instance where task interacts statistically significantly with V₂-UR-Status. As shown in the Results section, the statistically significant task-based difference between epenthetic [i] and lexical /i/ is found in the reading-in-a-frame, where [i] is statistically significantly shorter than /i/. In the other two tasks, however, [i] is marginally and non-significantly longer than /i/. In the absence of any disambiguating semantic clues to the vowel/zero contrast, [i] and /i/ are produced differently in the very task that encourages spelling-pronunciation.

Given that the vowel/zero contrast is not standardly represented in the orthography of BHA, do we conclude from this that the contrast is incompletely neutralised through [i]-epenthesis despite identical orthographic representation, but completely neutralised through [a]-epenthesis? Note that if orthography really has a very limited part to play in the phonetics of BHA, we should expect to find no quality-based differences that can be traced back to orthography.

An alternative interpretation is to conclude that the orthographic non-representation of the vowel/zero contrast induced a complete neutralisation effect in the case of 'a' but failed to do so in the case of 'i'. Again the quality-based difference is suspect. A recent psycholinguistic study by Bentin and Ibrahim (1996: 319) shows that when presenting words from an Arabic dialect in writing to speakers of that dialect, "[b]oth lexical decision and naming performance were inhibited". This is in contrast to the reaction by the same speakers to written words drawn from Standard Arabic. In other words, if lexical decision is inhibited for our subjects, we should expect the vowel/zero contrast to be completely neutralised for both 'a' and 'i'.

However, what is directly relevant to the genuineness question is that incomplete neutralisation seems to be cleared from being an orthographic artefact on either account. It is here complete neutralisation which seems suspect on orthography-related grounds. This is in contrast to what is commonly reported in the literature about the confounding potential of orthography. More specifically, orthography has only been available to blame when incomplete neutralisation is reported for a contrast that is represented orthographically (see e.g., Warner et al 2006, 2004).

At the same time, the durational difference between [i] and /i/ in the reading-in-a-frame task (but not in the other two tasks) where the semantics of the context is not neutral, seems to suggest that the distinction is to compensate for a degraded semantic context. Again, the disparate outcomes for 'a' and 'i' data raise doubt about claims that incomplete neutralisation is an artefact of the pragmatic context. Now we can see how the same compromised and artificial pragmatic context produces both complete and incomplete neutralisation. Interestingly, the naturalistic context in the other two tasks only produces complete neutralisation. Will this render incomplete neutralisation redundant, implying that complete neutralisation is actually an artefact of an unambiguous semantic and pragmatic context? This may agree to some extent with a suggestion made by Port and Crawford (1989) and defended later by Gafos (2006) that speakers can control how much contrast they 'want' to convey in neutralisation contexts. Nevertheless, it does not provide evidence for the genuineness argument. On the contrary, the results are more consistent with an alternative interpretation that does not appeal to the genuineness argument. This new interpretation states that the phonetics of neutralisation is variable, and lawfully so. In other words, both complete and incomplete effects are possible for the same context by the same speaker. This is the essence of the variability model I present in chapters six and seven.

As to the effect of stimulus manipulation in the paradigm, it seems that, by a statistical significance criterion, manipulation of stimuli has failed to have any effect on the acoustics of epenthetic and lexical vowels. V_2 s of words presented in a list where members of each minimal pair are kept as far apart as possible and those coming from words presented in a list where members of each minimal pair appear one immediately after the other are not produced differently along any of the acoustic measures of the study. The size of the raw mean differences also supports the above conclusion.

Just as in the case of task manipulation, that vowel production in this study emerges unaffected by manipulating speakers' awareness of the presence of minimal pairs is suspect. The literature on the phonetics of neutralisation provides us with a case where a similar manipulation has yielded a statistically significant effect. In Snoeren et al (2006), member-close lists of minimal pairs produced more differences between underlying and derived voiced obstruents in French than

members-apart lists. Again, just as in the case of task manipulation, the pattern might be related to the acoustics of vowels in BHA.

The box-plot data, however, suggest the possibility of stimulus-based F0 and intensity differences. The nature of these differences deserves further consideration. Ostentatiously drawing speakers' attention to minimal pairs, as in conditions 4-6, results in an increased variability within the intensity data for both vowel categories. But it has actually constrained variability within the F0 data for both vowel categories. The idea that minimal contrasts can constrain variation is not new. Although not always made explicit, it can be found under the rubric of a constrained phonetics because of the phonemic system of contrast in a language (see e.g., Campos-Astorkiza 2007; Lavoie 2002; Lindblom 1986; Manuel 1990; Tabain & Perrier 2007, 2005; Vaux & Samuels 2005). However, the increase in intensity variation is actually due to inter-speaker variation. For space reasons, I do not report nor discuss data from individuals in this thesis. Suffice it to observe here that the pattern seems to reflect speakers' different reactions to the highly contrast-promoting order of the stimulus set.

With regard to factor interactions, the i-intensity data show the only statistically significant interaction between stimuli and V₂-UR-Status. Specifically, [i] is statistically significantly more intense than /i/ in the members-apart list. The vowel/zero contrast is not enhanced in the members-close list. This finding seems to be at odds with Snoeren et al's (2006: 263) suggestion that incomplete neutralisation, especially where lexical confusability is involved, is "automatic in nature and reveal[s] the tight interdependency of productive and perceptual processes of lexical access in speakers/listeners' minds". The acoustic difference that BHA speakers produce in the naturalistic conditions does not show up in the conditions where lexical confusability is very likely—conditions where speakers produce members of minimal pairs in close succession. This result casts doubt on the claim that incomplete neutralisation is an automatic or intended strategy to counteract lexical ambiguity (cf. Gafos 2006; Port & Crawford 1989).

4.3.1.6 Interim Conclusion

This section has assessed the genuineness question from an experimental viewpoint by manipulating the naturalness of the experimental conditions. In the most naturalistic condition, subjects produce minimal pairs whose members are

orally elicited far apart. In the least naturalistic condition, subjects read out minimal pairs whose members appear one after the other. In this condition, the carrier sentence is a semantically neutral frame produced for all item words. The only clue to the meaning of the target member of each pair is a meta-linguistic descriptor specifying the morpho-syntactic category of the item in question. In between these two extremes, there are four conditions each with a different combination of the levels of stimuli, orthography, and pragmatic context.

Interestingly, the largest experimental differences holding between conditions 1 and 6 have mostly failed to show up in the phonetics of vowel/zero neutralisation in BHA. At least, this is wholly true of the neutralisation effect involving [a]-epenthesis. More specifically, [a]-epenthesis neutralises the vowel/zero contrast phonetically completely, even in the condition calling ostentatiously for a distinction to be produced.

In contrast, V_2 -UR-Status interacts statistically significantly with stimuli and with task for 'i'. These interactions suggest that [i]-epenthesis results in a phonetically incomplete neutralisation in conditions 1 and 6, respectively.⁶¹ What is noteworthy about these results is that vowel/zero neutralisation remains incomplete through [i]-epenthesis only in the two extreme conditions, the most naturalistic and the most unnaturalistic. This also means that [i]-epenthesis results in complete neutralisation in the remaining four conditions.

Of particular interest to the genuineness question is the results of 'a' and 'i' in condition 6. This is a condition that both promotes and demotes the realisation of contrast. This contradictory nature of the condition is due to the presentation of orthographically identical pair members in succession. In other words, the condition can potentially induce incomplete neutralisation when speakers are too

⁶¹ Compare the results of Condition 1 in this five-speaker experiment with those of the seven-speaker experiment reported in §4.2.1, where the only difference that reached statistical significance is a difference in intensity between epenthetic [a] and lexical /a/. On the basis of those results, I concluded that [a]-epenthesis resulted in incomplete neutralisation while [i]-epenthesis resulted in complete neutralisation. In contrast, the five-speaker data in this section display more differences between epenthetic [i] and lexical /i/ than between epenthetic [a] and lexical /a/. This inconsistency further supports the claim made in chapter two that if we choose to base our conclusions regarding the phonetics of neutralisation on NHST results, we will inevitably encounter cases where we find ourselves drawing mixed conclusions about the phonetics of neutralisation. The other interpretation of this is that the phonetics of neutralisation is variable between complete and incomplete effects. To choose between these competing views, we need a deep scrutiny of NHST procedures and outcomes. The next chapter offers just such a scrutiny.

aware of the presence of minimal pairs. Conversely, it can induce complete neutralisation when speakers pay more attention to the identical form of the words, or if they ignore or miss the lexical cue that the accompanying descriptors provide. Of course, individual speakers do not have to realise one and the same scenario. But in principle, they can.

The dilemma that the results of both vowel categories create so far is this: the same experimental make-up produces both complete and incomplete neutralisation; which do we choose to call an experimental artefact, and on what grounds? How naturalistic or otherwise an experimental condition is does not seem to qualify as a valid criterion. This will effectively render void the question of the genuineness of the experimental findings, at least as far as the experimental artefactuality argument is concerned. Next, I examine the genuineness question from a statistical angle.

4.3.2 Statistical Artefactuality

Compared to experimental artefactuality, statistical artefactuality has attracted relatively little attention within those circles disputing the genuineness of the neutralisation findings in the experimental literature. On those few occasions where statistical artefactuality is evoked, the statistical power of the experiment under scrutiny seems to be the main point of concern. Interestingly, it is complete neutralisation which mostly causes concern. Specifically, complete neutralisation has been disregarded as a side-effect of an experiment that is too low in power to detect the small differences that are common in neutralisation studies. To circumvent this problem, Warner et al (2006, 2004) call for carrying out production experiments using a large sample size. I take up the issue of statistical significance and sample size in the next chapter. But here I will focus on a much neglected aspect of the statistical artefactuality debate—one where both complete and incomplete neutralisation can equally be suspect, statistically speaking.

In particular, I will discuss the impact of pre-analysis on the statistical outcomes of different data analyses and subsequently on the inferences we draw based on such analyses. Limiting the discussion below to what is directly relevant to the phonetics of neutralisation, I will focus on two pre-analysis procedures that are commonly applied to data before statistical analysis. These are data aggregation and data pooling. For the purposes of this thesis, I shall make the following

distinction between these two procedures. Here, aggregation involves entering a single mean score for each individual or experimental unit in a given experimental condition. In contrast, pooling consists in entering multiple observations from an individual or a group of individuals as though they were independent observations in an experimental condition (see Leger & Didrichsons 1994). Put simply, pooling here does not involve any arithmetic calculation, whereas aggregation does.

Surveying the literature on the phonetics of neutralisation, one comes across various methods of data entry. Some rely exclusively on data aggregation; some on pooling; and some others combine the two procedures in the same analysis. Now it must be noted that not all methods of data entry are appropriate for the kind of statistical tests we run on the data we collect within a repeated-measures design. In (45), I have listed some of the common methods of data entry found in the relevant experimental literature.

(45) Common methods of data entry in the phonetics of neutralisation literature:

- i. Data averaged across repetitions and across items for each subject in an experimental condition
- ii. Data averaged across repetitions and across subjects for each item in an experimental condition
- iii. Data averaged for each item by each subject across repetitions only, then pooled for a given experimental condition
- iv. Data from each subject analysed separately
- v. Data pooled from all subjects, all items, and all repetitions, if applicable, for a given experimental condition

Obviously, methods (i) and (ii) above follow an aggregation procedure, whereas method (v) represents a complete pooling procedure. Method (iii), however, combines the two procedures by pooling aggregated data. Method (iv) is a special case of method (iii) if an individual's data are entered as an aggregate item pool; it is a special case of method (v), if raw data from the individual are entered.

Importantly, each one of these five methods assumes a different definition and leads to a different interpretation of what constitutes experimental units for statistical analyses. Specifically, experimental units are subjects according to (i), items according to (ii), each item by each subject according to (iii), all and only items of a single subject according to (iv), and the whole dataset according to (v). Defining experimental units is relevant to the validity of the outcome of Anovas and t-tests, which are, by far, the most common tests in the phonetics of neutralisation literature. For example, a basic assumption of these tests, commonly known as the independence assumption is very sensitive to the definition of experimental units—i.e., to the method of data entry. The independence assumption requires that observations in a given experimental condition within a repeated-measures design be independent of each other—i.e., come from different subjects (see below for more).⁶² Although Anovas and t-tests are often said to be robust against violations of a number of their basic assumptions (e.g., Howell 2002; Rietveld & van Hout 2005), many agree that a violation of the independence assumption is very serious for the validity of the outcome of these tests (e.g., Kenny & Judd 1986; Machlis et al 1985).

Importantly, Max and Onghena (1999) have shown how a violation of this assumption via a pooling procedure results in spurious statistical significance in the outcomes of Repeated-Measures Anova with a Huynh-Feldt correction, Manova, and a Mixed model analysis run on a hypothetical dataset.⁶³ However, when the same dataset are entered correctly, aggregated rather than pooled, none of these statistical tests display statistical significance. Likewise, on the basis of simulations, Jenkins (2002) concludes that the likelihood of finding spurious statistical significance, and thus committing a Type I error⁶⁴, is very high for pooled data as compared to aggregated data.

By definition, the pooling procedure leads to a violation of the independence assumption. This is commonly described as the ‘pooling fallacy’ (Machlis et al 1985). It follows then that the outcome of statistical tests on data entered using methods (iii), (iv), or (v) should be treated with caution. According to these

⁶² The technical definition of the independence assumption refers to the independence of the error components of the statistical model, but the simplified definition above is sufficient for our purposes (for more see Thoresen & Elashoff 1974).

⁶³ These results apply to t-tests, as well.

⁶⁴ In NHST, Type I error is said to occur when a true null hypothesis is rejected.

methods, the sample size (henceforth n), which enters the various equations necessary for the computation of the components of Anovas or t-tests, is erroneously over-estimated. What would be inappropriately treated as n according to these pooling methods does not reflect the true sample size. Specifically, n is equated with the product of the number of subjects and the number of items in (iii), with the number of multiple observations from an individual subject in (iv), and with the whole dataset in (v). This will grossly inflate the degrees of freedom (df) for the error in ANOVAs and t-tests, resulting in statistical significance being too easily reached. Experimental studies by Campos-Astorkiza (2008), Dmitrieva et al (2010), Gouskova and Hall (2009), Jassem and Richter (1989), Myers and Hansen (2005), Piroth and Janker (2004), Tabain and Perrier (2007), and Yu (2007), for example, report statistical analyses run on pooled data at high degrees of freedom. Charles-Luce (1985: 318) voices a similar complaint about inflated degrees of freedom in other studies to “near infinity”, which make it hard to compare statistical results that are pertinent to the same phenomenon but which are obtained from different studies.

Likewise, but perhaps less obviously, method (ii) incurs a violation of the independence assumption, even though it is essentially an aggregation procedure. To appreciate how that might be true, we first need to learn how and why method (i) is the only method listed in (45) that satisfies the independence assumption.

Method (i) takes each individual subject to represent one experimental unit. Put differently, for any given experimental condition in a repeated-measures design, each subject only contributes a single data point, averaged across repetitions, if any, and across items, if any. According to Max and Onghena (1999: 265), individual subjects are independent of each other in “the vast majority of research studies in the fields of speech, language, and hearing”. In contrast, items are usually nested within subjects, with each subject contributing a number of items. Accordingly, items coming from a single subject are not independent of each other.

Method (i) is commonly known as by-subject analysis. Here n equals the number of subjects. According to Raaijmakers (2003) and Raaijmakers et al (1999) (see also Locker et al 2007), this mode of analysis is the correct one for many repeated-measures datasets as found in language research, especially when item variation is experimentally controlled for (but see Rietveld & van Hout 2005 for a different

view). Among the experimental studies that offer by-subject analyses of neutralisation data are Barnes (2006), Charles-Luce (1985), Lahiri and Hankamer (1988), Slowiaczek and Dinnsen (1985), Snoeren et al (2006), and Warner et al (2006, 2004).

In contrast, method (ii), known as by-item analysis, averages across subjects and across repetitions. Since items are the experimental units in this analysis, n is equated with the number of items analysed. Originally, this method has been proposed to resolve what is now known as the ‘language-as-fixed-effect fallacy’ (see Clark 1973 for details). The outcome of a by-item analysis together with the outcome of a by-subject analysis are meant to provide the necessary components for the calculation of $\min F'$, which can be used for hypothesis testing.⁶⁵ In other words, a by-item analysis is not meant for direct hypothesis testing, contra what many studies these days seem to suggest. Raaijmakers (2003: 146) evaluating such a development, finds no “statistical rationale for such a procedure”. Moreover, according to Rietveld and van Hout (2005: 169), since by-item analyses do not take into account correlations within subjects, they “tend to give incorrect inferences”. This follows from their violating the independence assumption. Experimental studies by Mitleb (1981), Port and O’Dell (1985), Snoeren et al (2006), Tieszen (1997), Warner et al (2006, 2004), for example, contain by-item analyses of neutralisation data.

What is directly relevant to the genuineness question from the discussion above is that using different methods of data entry, which each entails a different definition and interpretation of experimental units in the statistical analysis, can give quite different statistical outcomes. However, it will be inadequate to explore statistical artefactuality solely by comparing the statistical outcomes of the different studies mentioned above which follow different methods of data entry. There is a multitude of design-related differences among those studies, which may greatly reduce comparability among them. For example, mean differences vary widely in those studies. Moreover, the magnitude of neutralisation effects and data variation

⁶⁵ $\min F'$ is actually the lower bound of F' , which is originally proposed for evaluating hypotheses involving both subjects and materials. But due to its ease of calculation, $\min F'$ is often used instead.

On how to calculate the degrees of freedom associated with $\min F'$, see Clark (1973).

does not show up to a similar degree. Realistically, these differences will confound any effect that might be attributed to the methodology of data entry.

To isolate the contribution of pre-analysis procedures, we need to control for the various confounding design-related variables. One way of doing this is through simulation—but simulated studies have already been reported and referred to above to illustrate the point being made. For more on this option, see Jenkins (2002) and Max and Onghena (1999). Another way, which I present below, is to submit real neutralisation data to different pre-analysis procedures, thus testing the effect of data-entry methodology on statistical outcomes and subsequently on our conclusions regarding statistical artefactuality and regarding the phonetics of neutralisation in general.

Specifically, I re-analyse Turkish and Polish devoicing data from Kopkallı (1993) and from Tieszen (1997), respectively. Unlike many studies in the literature, these two studies provide detailed descriptive statistics including an adequate breakdown of mean scores by subjects, items, and conditions. The choice of the specific methods of data entry I try here is not arbitrary; it is rather constrained by the form of the available data. For example, Kopkallı (1993) does not give raw data. Thus, a re-analysis using method (v) is not possible for Turkish. Similarly, Tieszen (1997) does not include multiple repetitions of each item by each speaker. Thus, a re-analysis using method (iii) is not possible for Polish.

Kopkallı (1993) (henceforth Kopkallı) fitted multi-factor mixed models to data from each of the five subjects in her experiment. These factors include the underlying voicing of the target word-final consonant, four temporal parameters (i.e., preceding vowel, consonant, pulsing, and aspiration), two contexts (sentential, and in isolation), three places of articulation (bilabial, dental, and velar), and familiarity of test words. In addition to these fixed factors, the model includes several inter-factor interaction terms, and a random effect of words. A χ^2 test is used to evaluate the contribution of these factors. Kopkallı explains that she resorted to individual data analyses to simplify the models, which otherwise would have required larger computer memory than was available to her. In none of the models fitted is underlying voicing statistically significant. However, underlying voicing statistically significantly interacts with temporal parameter for two out of

the five subjects, implying that for some parameters, underlying voiced and voiceless consonants are realised differently by these speakers. Nevertheless, Kopkallı decided not to investigate this any further and concluded that final devoicing in Turkish is completely neutralising.

To re-analyse these Turkish data, I submitted them to an aggregation procedure (method i) and to a pooling procedure (method iii). Table 4-18 summaries mean paired differences (\bar{X}_{PD}) and their SD_{PD} values, according to methods (i) and (iii).



Table 4-18: Descriptive statistics for Turkish final devoicing including mean paired differences and their standard deviations as calculated using a by-subject aggregation method (i) and a pooling method (iii); data come from Kopkallı (1993).

As is clear from the table above, aggregation and pooling yield different variation estimates in the form of SD_{PD} . In particular, aggregation tends to under-estimate variability within the data (see Leger & Didrichsons 1994), while pooling over-estimates the degrees of freedom for statistical inferential testing. Importantly, however, mean differences are unaffected by these pre-analysis procedures.

A series of paired t-tests run on the by-subject aggregated data (method i) with a Bonferroni correction reveal statistically significant differences along two of the four parameters of the study: vowel duration [$t(4) = 5.5$; $p = .005$] and closure pulsing duration [$t(4) = 5.85$; $p = .004$].

The outcome of the paired t-tests run on the pooled data (method iii) with a Bonferroni correction is not the same as above. Here three out of the four parameters of the study show statistically significant differences. These are vowel duration [$t(29) = 3.5$; $p = .001$], consonant duration [$t(29) = -3.5$; $p = .002$], and closure pulsing duration [$t(29) = 3.3$; $p = .002$]. Table 4-19 gives the results of the paired t-tests on the aggregated and pooled data. Note that without adjusting for

multiple testing⁶⁶, the outcome of the pooling procedure would have been even more dramatically different from the outcome of the aggregation procedure and indeed from Kopkallı's own individuals' data analyses. Without adjusting for multiple testing, the pooling procedure finds statistical significance along all the temporal parameters in the study.

	Aggregation (method i)		Pooling (method iii)	
	t df=4	P	t df=29	p
V-Duration	5.5	.005*	3.5	.001*
C-Duration	-2.66	.057	-3.48	.002*
Pulsing-Duration	5.85	.004*	3.34	.002*
Asp-Duration	1.78	.150	2.11	.044

*statistically significant at the Bonferroni-adjusted alpha at 5%

Table 4-19: Results of two-tailed paired t-tests for Turkish final devoicing

Now the important question that is pressing us for an answer is this. Which effect is genuine? Is Kopkallı's conclusion of complete neutralisation, which is based on individual data analyses, a statistical artefact? Is our new conclusion of incomplete neutralisation based on a by-subject analysis genuine? Or is a conclusion of incomplete neutralisation that a pooled aggregate analysis supports a statistical artefact? Of course an obvious way to pursue these questions further is to scrutinise the statistical validity and soundness of the outcome of the NHST procedures as specifically applied to the Turkish data. In the next chapter, I take up the criterion of statistical significance more generally. However, for now, we need to reflect on the implications of these results for the genuineness argument—the topic of this section.

COPYRIGHT MATERIAL

⁶⁶ See the discussion in §4.2.1.2.3.3 on multiple testing.

COPYRIGHT MATERIAL

Now concluding that voicing neutralisation in Turkish is phonetically incomplete will mean that incomplete neutralisation can occur despite an inauspicious orthographic representation of neutralisation. Such an outcome indicates that incomplete neutralisation is not necessarily a spelling pronunciation or an experimental artefact, at least as far as the Turkish data at hand are concerned. However, the verdict for the question of experimental artefactuality seems to lie with statistics.

Importantly, the statistical re-analysis of the Turkish data above suggests that statistical artefactuality can be affixed not only to complete neutralisation, as has been usually the case, but also to incomplete neutralisation. It is fortuitous that a single case, the Turkish data, which illustrates the duality of the experimental artefactuality above, should furnish us with an opportunity to dissect the statistical artefactuality issue, thus vitiating the genuineness argument from both experimental and statistical perspectives.

Now consider Tieszen's (1997) (henceforth Tieszen) neutralisation data and conclusions regarding final devoicing in Polish. Tieszen explores the phonetics of final devoicing in three dialects of Polish: Warsaw, Bydgoszcz, and Kraków. On the basis of statistical analyses run on data from five native speakers from each dialect producing nine minimal pairs, Tieszen concludes that neutralisation is phonetically incomplete in Warsaw, but complete in Kraków, with Bydgoszcz falling in between, as a transitional dialect. Tieszen's analysis applies a by-item aggregation

procedure. Tieszen has investigated three acoustic parameters. These are vowel duration, closure duration, and closure pulsing duration. Tests items are embedded into two carrier contexts: ___#C (-voice) and ___#vowel. I limit my re-analysis here to the consonant-initial environment, following a hint from Tieszen that the phonological environment for final devoicing in Kraków does not include a vowel-initial following context. Also, for the sake of simplicity, I do not re-analyse the data from the transitional dialect of Bydgoszcz.

Tieszen reports that, in Warsaw Polish, a mean durational difference in closure pulsing between underlyingly voiced and voiceless stops of ≈ 20 ms is statistically significant: $[t(8) = 4.13; p < .05]$. No statistically significant differences are found in Kraków Polish.

Here, I present statistical analyses of these Polish data, having submitted them to an aggregation procedure (method i) and a pooling procedure (method v). Table 4-20 and Table 4-21 give mean paired differences (\bar{X}_{PD}) and their SD_{PD} values for the data from Warsaw Polish and Kraków Polish, respectively.⁶⁷

COPYRIGHT MATERIAL

Table 4-20: Descriptive statistics for Warsaw-Polish final devoicing including mean paired differences and their standard deviations as calculated using a by-subject aggregation method (i) and a pooling method (v); data come from Tieszen (1997).

COPYRIGHT MATERIAL

Table 4-21: Descriptive statistics for Kraków-Polish final devoicing including mean paired differences and their standard deviations as calculated using a by-subject aggregation method (i) and a pooling method (v); data come from Tieszen (1997).

⁶⁷ There is a very slight difference between some of the mean figures reported in Tieszen (1997) and the calculated values I summarise in Table 4-20 and Table 4-21. This difference should be seen as a rounding error.

Just as we observe for the Turkish data, aggregation and pooling procedures result in different values of SD_{PD} , but have no effect on mean paired differences. While aggregation seems to under-estimate variability within the data, pooling actually over-estimates the degrees of freedom for the statistical inferential testing that I report below.

A series of paired t-tests run on the by-subject aggregated data (method i) with a Bonferroni correction found no statistically significant differences along any of the three temporal parameters for either dialect. In contrast, the pooling procedure reveals a statistically significant difference in closure pulsing duration for both dialects: Warsaw [$t(44) = 5.2$; $p < .001$], Kraków [$t(44) = 2.56$; $p = .014$]. Table 4-22 and Table 4-23 present the results of the paired t-tests run on the aggregated and pooled data for Warsaw and Kraków, respectively.

	Aggregation (method i)		Pooling (method v)	
	t df=4	P	t df=44	p
V-Duration	2.1	.104	2.2	.035
C-Duration	-3.04	.039	-1.77	.083
Pulsing-Duration	3.78	.020	5.23	.000*

*statistically significant at the Bonferroni-adjusted alpha at 5%

Table 4-22: Results of two-tailed paired t-tests for Warsaw-Polish final devoicing

	Aggregation (method i)		Pooling (method v)	
	t df=4	P	t df=44	p
V-Duration	1.95	.124	.97	.336
C-Duration	.67	.542	.53	.598
Pulsing-Duration	1.29	.268	2.56	.014*

*statistically significant at the Bonferroni-adjusted alpha at 5%

Table 4-23: Results of two-tailed paired t-tests for Kraków-Polish final devoicing

Consider next the conclusion we get on the basis of a minF' analysis of these data. As I have pointed out earlier, the by-item F-ratio that Tieszen reports should not be used for significance testing. Instead, a by-subject analysis or a minF' ratio should

be more appropriate for that purpose (see Raaijmakers 2003; Raaijmakers et al 1999). Interestingly, the minF' analysis fails to find statistical significance along any of the parameters investigated for either dialect. See Table 4-24 and Table 4-25 below. As we can see from the tables, statistical significance derived on the basis of minF' analyses is similar to that derived on the basis of by-subject analyses (method i) after correcting for multiple testing.

COPYRIGHT MATERIAL

Table 4-24: Results of Repeated-Measures Anovas for Warsaw-Polish final devoicing

COPYRIGHT MATERIAL

Table 4-25: Results of Repeated-Measures Anovas for Kraków-Polish final devoicing

Just as what we experience with the Turkish case above, we seem to arrive at different interpretations of the Polish data by adopting different pre-analysis procedures. Specifically, Tieszen's (1997) conclusion based on a by-item aggregation procedure is different from a conclusion one might draw on the basis of a by-subject aggregation procedure and from a conclusion resting on a pooled-data analysis. According to Tieszen, voicing neutralisation is phonetically incomplete in Warsaw Polish, but complete in Kraków Polish. By both a by-subject

analysis and a minF' analysis, however, neutralisation is complete in both dialects.

COPYRIGHT MATERIAL

The outcome of this re-analysis of the Polish data in Tieszen (1997) lends further support to the conclusions I presented above on the basis of my re-analysis of the Turkish data. An effect that has been reported, cited, and re-cited in the literature as a case of complete neutralisation could have just as easily been reported, cited, and re-cited as a case representing incomplete neutralisation. Conversely, an effect classified as incomplete neutralisation on the basis of some statistical testing could have been classified otherwise on the basis of some slightly different statistical testing. Note that statistical artefactuality is not necessarily tied to statistical power here. Put differently, claims that complete neutralisation is a statistical artefact are only telling part of the story. Incomplete neutralisation can just as equally possibly be nothing more than an inflated statistic. In the next chapter, I argue that incomplete neutralisation is more prevalent on purely statistical grounds. This prevalence is actually something we observe in the literature on the phonetics of neutralisation (see chapter two and Table 4-14 above).

4.4 Conclusion

In this chapter, I explored the phonetic completeness of vowel/zero neutralisation in BHA. I presented acoustic and perceptual data from native speakers of BHA. The acoustic data come from a simple experimental design with one independent variable—the vowel/zero underlying contrast that vowel epenthesis supposedly neutralises. Results of the production experiment reveal a curious pattern of dis-correlation between the phonetics and phonology of neutralisation. Epenthetic [a] and lexical /a/, which behave the same in the phonology, are distinct in the phonetics in that epenthetic [a] is statistically significantly more intense than

lexical /a/. Conversely, epenthetic [i] and lexical /i/, which behave differently in the phonology, are phonetically identical. No less curious, though, is the apparent utilisation of a contrastively inactive acoustic cue for the purposes of preserving a contrast that seems to be completely neutralised in the phonology. Perception-wise, [a] and /a/ are discriminated less accurately than are [i] and /i/. I discussed the implications of these unexpected results for the laboratory tradition in the study of the phonetics of neutralisation.

I then discussed the genuineness question from an experimental and statistical point of view. Experimentally, my arguments benefit from insights emerging from an experimental paradigm that has manipulated important variables claimed to exert an influence on the phonetics of neutralisation. These variables include orthography, pragmatic context, and the presence of minimal pairs in the stimulus list. Statistically, I re-analysed real neutralisation data from Turkish and Polish, applying different pre-analysis procedures to these data and deriving mixed conclusions regarding the phonetics of neutralisation.

My main conclusion concerning the genuineness argument is that it is equivocal and uninformative. It is equivocal because doubting the genuineness of a set of data on the grounds that the observed effect is an experimental artefact brought about by the specifics of the experiment design can equally well apply to both complete and incomplete neutralisation. Put differently, as much as this logic casts doubt on either effect, it legitimises both. Conversely, dismissing the observed effect as a statistical artefact can be defended for both complete and incomplete neutralisation.

The genuineness argument is uninformative because it is essentially a filtering criterion. It encourages censorship among researchers and commentators. Instead of thinking constructively about the theoretical and practical implications of a set of findings, researchers will become more concerned with the reliability and reproducibility of the findings. The genuineness argument has, until very recently, discouraged the integration of the phonetic findings of neutralisation studies into theories of phonetics and phonology. There is a real need for an informed reassessment of the set of criteria we use to draw conclusions regarding the phonetics of neutralisation in terms of both the completeness question and the genuineness question. The next chapter attempts to offer just that.

5 Characterising and Quantifying the Phonetics of Neutralisation

5.1 Introduction

A central goal of this thesis is to increase our understanding of the phonetics of neutralisation and the variability therein. The discussion of the phonetics of vowel/zero neutralisation in BHA presented in chapter four highlights the need to re-consider (1) our use of statistical significance as a labelling criterion to qualitatively describe the phonetics of neutralisation and (2) the parametric measures we use to quantify the phonetics of neutralisation. I touched on these issues when I looked into both descriptive and inferential statistics in the previous chapter. I elaborate on them here.

Of special importance to the current chapter is the issue of variability. As I have suggested earlier, there are two types of variability that we need to consider in our study of the phonetics of neutralisation: an inherent quantitative variability and an acquired qualitative variability. The latter follows from our drawing a qualitative distinction between phonetically complete and incomplete neutralisation. Obviously, this qualitative distinction is based on a set of criteria that have been commonly used to classify a phonetic effect as ‘complete neutralisation’ or as ‘incomplete neutralisation’. I described and illustrated qualitative variability in chapter two. Here I will focus on the labelling criteria in the phonetics of neutralisation.

To decide on a label to describe the phonetics of a certain neutralisation, researchers have traditionally resorted to a statistical assessment procedure of the significance of the phonetic difference found, its directionality, and its perceptibility.

Considering the current practices within our field, we find that statistical significance has become a requirement for establishing the existence or otherwise of a difference, its directionality, and perceptual relevance. The definition of each of these criteria involves an appeal to statistical significance. Therefore, I focus on statistical significance in my discussion of the labelling criteria in §5.2.

To keep the discussion in perspective, I consider how statistical significance relates to practical (or equivalently in our situation linguistic) significance. Practical significance can potentially become a sensitive issue for the phonetics of neutralisation, especially for the notion of incomplete neutralisation.⁶⁸ In particular, I make the claim that statistical significance does not necessarily imply linguistic significance. I show how this claim reflects a distinction made and debated some time ago in a few fields of research that rely for their scientific conclusions on NHST. Although the distinction is hardly recognised in our field, researchers occasionally appeal to it as a last-resort option, especially when they are confronted with an unexpected result that defies linguistic explanation. I conclude the discussion by observing that conclusive evidence for or against the complete-incomplete distinction is not likely to emerge from the set of labelling criteria, at least as currently (mis)used.

With regard to the summary measures we use to quantify the phonetics of neutralisation, I show that the standardly used measures of central tendency (the mean) and variability (SD or RSD) are unintuitive and so closely tied to the numerical values of the measurement scale as to potentially undermine any robust estimation of the underlying central location and variability. For example, our calculation of the mean can be very easily contaminated by outliers or extreme values in a dataset. Outliers can also affect the SD, but to a lesser extent. RSD, however, can be entirely dependent on the numerical value of the mean.

I evaluate these measures and propose an alternative that is both more intuitive and cognitively plausible. Specifically, I suggest that the mode, rather than the arithmetic mean, be used to measure central tendency. The mode reflects better the intuitive notion of average. Likewise, the variability measure I propose (VFI)

⁶⁸ The vanishingly small differences found in reports of incomplete neutralisation underline the need for establishing that the differences are not only statistically significant but also practically significant.

relates the frequency of the modal interval to the range. The frequency of the modal interval is independent of the numerical value of the measurement scale, and more in line with frequency-based Bayesian reasoning (Gigerenzer & Hoffrage 1995). Similarly, the range captures the intuitive notion of variation. One particular issue I dwell on is how to find the mode for continuous data, the most common type of phonetic data. I suggest that phonetic data are more appropriately examined as intervals rather than as single points. In the next two chapters I elaborate on this suggestion and propose a binning algorithm utilising the familiar psycho-physical notion of just noticeable difference (jnd).

The rest of the chapter proceeds as follows. In §5.2, I scrutinise the labelling criteria in the literature as applied to the phonetics of neutralisation. I first present a brief overview of these criteria in §5.2.1. I then discuss in §5.2.2 statistical significance, the single most important criterion in the analysis of neutralisation data. In §5.3.1, I evaluate the parametric measures of central tendency and dispersion that are commonly used to quantify the phonetics of neutralisation. I then introduce an alternative and highlight its intuitiveness in §5.3.2. I conclude the chapter in §5.4.

5.2 Characterising the Phonetics of Neutralisation: Labelling Criteria

5.2.1 Overview

Reviewing the literature on the phonetics of neutralisation, we observe that the most defining feature of incomplete neutralisation is not only that there is an acoustic difference between two sounds that are said to have neutralised, but also (and more importantly) that the difference is (1) statistically significant, (2) in the expected direction, and (3) usually perceptually relevant⁶⁹. From these, we can infer the criteria that have been standardly used to label a phonetic effect as

⁶⁹ There is widespread agreement that a difference in perceptibility necessarily implies a difference in production (see e.g., Dinnsen 1985; Jongman 2004). By contrast, there is no consensus on whether or not a difference in production should be at all hearable (see e.g., Brockhaus 1995; Labov 1994; Manaster Ramer 1996a, 1996b). For reasons of space I do not elaborate on the production-perception relation here. But see my discussion on this issue in §4.2.3 and §5.2.2.

'complete' or as 'incomplete'. Apparently, these are (1) statistical significance, (2) directionality, and (3) perceptibility.

However, this characterisation of incomplete neutralisation is not precise. For example, it overlooks the important point that both directionality and perceptibility are conventionally defined in terms of statistical significance. The standard practice we follow in drawing any conclusions regarding directionality and perceptibility, or indeed any other issue we submit to statistical inference, is this. To conclude that a difference is 'going in the expected direction' or is 'perceptible', researchers find themselves compelled by the force of tradition and accepted practices to demonstrate that the difference is statistically significantly going in that direction and that it is statistically significantly perceptible. Differences, be it in directionality or perceptibility, that do not reach statistical significance, or just fall short of it, are usually ignored, dismissed as random errors, or described as mere 'trends' that are incapable of supporting any conclusive findings. In short, statistical significance seems to be a necessary component of the definition of both directionality and perceptibility.

Perhaps a better characterisation of the situation is to explicitly include statistical significance as a necessary component (perhaps the only necessary component) of the definition of all the labelling criteria. Granted, this would involve assessing the statistical significance of a phonetic difference in terms of (1) existence, (2) directionality, and (3) perceptibility. Underlying these criteria are the questions in (46).

(46) Questions underlying the labelling criteria

Q1: Is there a statistically significant difference?

Q2: Is it statistically significantly in the expected direction?

Q3: Is it statistically significantly perceptible?

Now looking at these questions, we may note that directionality (Q2) and perceptibility (Q3) are appended to what is essentially Q1, as shown in (47).

(47) Questions underlying the labelling criteria modified

Q1: *Is there a statistically significant difference?*

Q2: *Is there a statistically significant difference in directionality?*

Q3: *Is there a statistically significant difference in perceptibility?*

So have we objected to the imprecision of the characterisation at the beginning of the section, only to come up with a characterisation that suffers from redundancy? Well, it is instructive to note that the imprecision in the characterisation at the beginning of the section is potentially misleading. For example, it seems to be erroneously suggesting that directionality and perceptibility can be established independently of statistical significance, which is not the case as yet.

On the other hand, the overt redundancy in (47) is useful because it allows us to see that statistical significance is actually the most relevant component of the definition of each of these criteria. Thus, a scrutiny of what statistical significance means in general seems adequate for an overall qualitative description of the phonetics of neutralisation.

5.2.2 Statistical Significance versus Practical Significance

Statistical significance is a basic ingredient of the definition of each of the criteria that we standardly use to distinguish between neutralisation effects that are phonetically complete and those that are phonetically incomplete. The distinction rests on whether or not a phonetic difference reaches statistical significance. By convention, the pre-specified significance level used for the decision to reject or not to reject the relevant null hypothesis (H_0) is set at .05.

When it comes to qualitatively describing and interpreting findings in light of how closely the achieved p value approximates the .05 level, researchers employ a limited set of phrases such as ‘is not significant’, ‘is just significant’, ‘has approached significance’, ‘has just missed significance’, and, obviously, ‘is significant’, and ‘is highly significant’. Here, the smaller the p value, the higher the statistical significance is attributed to it.

I have no doubt that researchers in our field take the p value as at least some kind of measure of statistical significance. Yet I am not as certain that all of them conduct their interpretation of the results of statistical tests within the confines of

what statistical significance really means.⁷⁰ A similar concern has been voiced by researchers from a wide variety of research backgrounds such as psychology and education (e.g., Bakan 1966; Carver 1978; Cohen 1994; Daniel 1998; Gigerenzer et al 2004; Gliner et al 2002; Haller & Krauss 2002; Hunter 1997; Scarr 1997; Schmidt & Hunter 1997; Thompson 1998a, 1998b), economics and marketing (e.g., Hubbard & Armstrong 2006; McCloskey & Ziliak 1996; Sawyer & Peter 1983), forecasting (e.g., Armstrong 2007a, 2007b), and medical research (e.g., Altman et al 2000). In his “*The earth is round (p < .05)*” article, Cohen (1994: 997) complains about a “near-universal misinterpretation” of statistical significance. A survey study by Haller and Krauss (2002) shows that 80% of the sampled professors and lecturers teaching statistics to psychology students in six German universities gave wrong answers about the meaning of statistical significance in terms of ($p = .01$). These wrong answers come from a list of some of the most common misconceptions of statistical significance, given in (48).

(48) Common misconceptions of $p = .01$ (adapted from Haller & Krauss 2002: 5)

COPYRIGHT
MATERIAL

⁷⁰ Note, for example, that Hubbard and Lindsay (2008: 71) take the fact “that p values continue to saturate empirical work [...] as *prima facie* testimony that most psychology (and other) scholars [...] remain unaware of many of the reasons why this index is a defective measure of evidence”. More generally, Tryon (1998: 796), lamenting the contemporary widespread confusion over statistical significance, writes that “the fact that statistical experts and investigators publishing in the best journals cannot consistently interpret the results of these analyses is extremely disturbing. Seventy-two years of education have resulted in minuscule, if any, progress toward correcting this situation”.

To appreciate the falsehood of the different interpretations in (48), let us start with the correct interpretation of $p = .01$. Statistically speaking, a p value that a null-hypothesis significance test yields is *not* the probability that the null hypothesis is true. Rather, it only indicates “the probability of getting a test statistic which is as extreme, or more extreme, than the one observed, assuming that H_0 is true” (Chatfield 1995: 249). In other words, probability here pertains to the observed and some more extreme unobserved data but not to the null or alternative hypotheses. As such, it says nothing about the truth or otherwise of these hypotheses. The misinterpretations in (48) commonly originate from confusing the answers to these questions: (1) what is the probability (p) of getting the observed data (D) given that H_0 is true (i.e., $p(D|H_0)$) and (2) what is the probability of H_0 being true given the observed data (i.e., $p(H_0|D)$)? The p -value only provides an answer to question (1). However, many researchers wrongly take it as an answer to the question they need an answer to—question (2) (see e.g., Tryon 2001; Carver 1978; Cohen 1994). A researcher not steeped in NHST will probably tend to underplay the practical and theoretical consequences of this confusion (cf. Berger & Sellke 1987). However, Carver (1978: 384), along with others, tells us that taking $p(D|H_0)$ and $p(H_0|D)$, knowingly or not, to be interchangeable can be an egregious mistake. He writes:

What is the probability of obtaining a dead person (label this part D) given that the person was hanged (label this part H); this is, in symbol form, what is $p(D|H)$? Obviously, it will be very high, perhaps 0.97 or higher. Now, let us reverse the question. What is the probability that a person has been hanged (H), given that the person is dead (D); that is, what is $p(H|D)$? This time the probability will undoubtedly be very low, perhaps 0.01 or lower. No one would be likely to make the mistake of substituting the first estimate (0.97) for the second (0.01); that is, to accept 0.97 as the probability that a person has been hanged given that the person is dead. Even though this seems to be an unlikely mistake, it is exactly the kind of mistake that is made with interpretations of statistical significance testing --- by analogy, calculated estimates of $p(H|D)$ are interpreted as if they were estimates of $p(D|H)$, when they are clearly not the same.

In 2006, Hubbard and Armstrong published a paper entitled “Why we don’t really know what statistical significance means”. They reiterated a conclusion that many before them had reached:

[T]he principal reason that researchers cannot accurately define what is meant by statistical significance is that many statistics and methodology textbooks are similarly confused over the exact meaning of this concept (p. 115).⁷¹

Actually, the elusive nature of the concept of statistical significance has been but one of the mildest forms of the on-going attack on the whole procedure of NHST (see e.g., Armstrong 2007a, 2007b; Bakan 1966; Berger & Selke 1987; Carver 1978; Cohen 1994; Jones & Tukey 2000; Tyron 1998). A more serious criticism of NHST is that statistical significance can simply be a question of sample size.⁷² Consider the following excerpts from a few disillusioned NHST practitioners:

Virtually any study can be made to show significant results if one uses enough subjects regardless of how nonsensical the content may be (Hays 1963: 326).

Indeed, given enough subjects and middling hypotheses, some significant but trivial findings are a certainty (Robey 2004: 311).

Hence, because type I errors cannot occur, statistically significant results are assured if large enough samples are used (Kirk 2007: 1636).

Statistical testing becomes a tautological search for enough participants to achieve statistical significance. If we fail to reject, it is only because we've been too lazy to drag in enough participants (Thompson 1998b: 799).

[S]mall differences of no real interest can be statistically significant with large sample sizes, whereas clinically important effects may be statistically non-significant only because the number of subjects studied was small (Altman et al 2000: 17).

⁷¹ The real roots of the problem, as argued by Hubbard and Armstrong (2006) and Gigerenzer et al (2004), among others, can be traced back to an unfortunate and illegitimate merging of Fisher's 'evidential' p value and the Neyman-Pearson α (Type I error rate). These two measures of 'statistical significance' have different functions. Their integration into one model has not been endorsed by their originators (see e.g., Hubbard & Bayarri 2003; Kirk 2007; Sterne & Davey Smith 2001). Very crudely, Fisher's p is meant to evaluate evidence against the hypothesis that is put forth to be nullified within a single study, whereas the Neyman-Pearson α is the probability of rejecting a true hypothesis in repeated studies. The Neyman-Pearson α has to be pre-specified together with β , the probability of accepting a false hypothesis (also known as Type-II error rate). In other words, the NHST paradigm that we currently apply is actually a hybrid that is, in Gigerenzer et al's (2004: 400) words, "a mishmash that does not exist in statistics proper". Tracing the history of this confusion along with its subsequent repercussions lies beyond the scope of this thesis (for more, see Gigerenzer et al 2004; Harlow et al 1997; Huberty 1993; Menon 1993; McClure & Suen 1994).

⁷² There are researchers who might occasionally acknowledge that statistical significance is sensitive to sample size but maintain, nonetheless, that the correlation is much less pronounced for 'true' null hypotheses. They take comfort in the fact that at least the probability of obtaining statistical significance under a true null hypothesis never goes to 1.00 (e.g., Nickerson 2000). But see the discussion below on the 'truth' of null hypotheses.

The close correlation between statistical significance and sample size can be demonstrated using simulated experiments. For example, Marsh et al (2008) report three virtual studies involving two independent groups with a mean difference of 5 between the two groups and an SD of 15 in each group. The p -value that is associated with the independent-samples t -test becomes increasingly smaller as the sample size gets larger. This is illustrated in Table 5-1. Our interpretation of a mean difference of 5 swings from being statistically non-significant when $n=10$ to being statistically highly significant when $n=100$.

COPYRIGHT MATERIAL

Table 5-1: Statistical significance and sample size: independent-samples t -test (based on Marsh et al 2008)

In Repeated-Measures experiments, *the* experimental design in neutralisation studies, statistical significance can be reached even with a far smaller number of subjects than in the Marsh et al study for the same mean difference. This can be demonstrated by simulated⁷³ experiments involving a random sampling of subjects to be tested in two conditions. The parameters of the population from which the sample is taken are as follows⁷⁴: μ in condition one =105; μ in condition two =100; $\sigma = 15$; and $\rho = .5$. Results of paired t -tests appear in Table 5-2.

	Study 1	Study 2	Study 3	Study 4	Study 5
N	10	15	20	25	30
t	1.37	2.31	2.66	2.81	2.97
p	.20	.036	.015	.009	.005

Note: μ in condition one =105; μ in condition two =100; $\sigma = 15$; $\rho = .5$

Table 5-2: Statistical significance and sample size: two-tailed paired-sample t -tests

⁷³ I used the Rice Virtual Lab in Statistics (<http://onlinestatbook.com/rvls.html>) to run the simulations.

⁷⁴ To avoid confusion with the p -value, I use the full form **rho** rather than the more appropriate ρ symbol.

We can also form an initial impression of the close correlation between statistical significance and sample size by comparing the percentages of experiments yielding statistical significance in different size-conditions. Consider the percent figures in Table 5-3. These are generated from 30000 simulations of experiments with a repeated-measures design where the samples are randomly drawn from a population with the following parameters⁷⁵: μ in condition one =60; μ in condition two =70; σ =8; and ρ =.5. The percentage of the experiments that yield statistical significance increases from 55% when n =5 to almost 100% when n =20. The exact figures are given in Table 5-3.

		Number of studies yielding $p < .05$	Number of studies yielding $p > .05$	% Studies showing statistical significance
Number of subjects	N=5	2797	2203	55%
	N=10	4667	333	93%
	N=15	4966	34	99%
	N=20	4998	2	100%
	N=25	5000	0	100%
	N=30	5000	0	100%

Note: μ in condition 1 =60; μ in condition 2 =70; σ =8; ρ =.5

Table 5-3: Sample size and percentage of studies yielding statistical significance out of 5000 simulated experiments in each size-condition: Repeated-Measures Design

The consequences for the current statistical treatment of the phonetics of neutralisation are unsettling indeed. Interpreting the outcome of NHST relative to the phonetics of neutralisation, researchers seem to unanimously endorse but one strategy: statistical significance is interpreted as incomplete neutralisation, statistical non-significance as complete neutralisation.⁷⁶ In light of the preceding shortcomings of the NHST procedure, one can see where this strategy would lead us: incomplete neutralisation would be a certainty while complete neutralisation, being counter to a certainty, would be a logical impossibility.

⁷⁵ The same pattern is observed even with smaller effects and within different statistical models as well. For example, Baayen (2008), using simulated data, demonstrates how a regression model that explains only 1% of variation can be, nonetheless, statistically significant when n =1000.

⁷⁶ Note that under those rare circumstances where deep misgivings are expressed about a particular NHST result, it is the validity of the result that is usually questioned, not the viability of the strategy of equating statistical significance with incomplete neutralisation. A researcher who is unprepared to accept incomplete neutralisation may protest that statistical significance is spurious in the context where it has been found.

But the impossibility of logically deriving complete neutralisation also stems from its being construed of in the neutralisation literature as a null hypothesis of no difference. The dominant view of H_0 in our field is that it is a hypothesis of no difference between two acoustic events along some measured acoustic parameters. There are at least two interpretations of a no-difference hypothesis: (1) there is no difference between A and B, and (2) there is no statistically significant difference between A and B.⁷⁷ In symbol notations, the first interpretation reads as ($A - B = 0$) while the second as ($A - B = \text{a test statistic with } p \geq .05$). Now to say that there are no differences between two events (i.e., $A - B = 0$) is never going to be true given the very large number of decimal places where the difference can show up (Tukey 1991; Jones & Tukey 2000). This is why, logically, any no-difference hypothesis is always going to be false. Similarly, to say that there are differences that are statistically non-significant (i.e., $A - B = \text{a test statistic with } p \geq .05$) can seldom be true in a study that uses a large number of participants. Actually, many⁷⁸ statisticians and empirical researchers hold the view that “null hypotheses are never true except those we construct for Monte Carlo tests of statistical procedures” (Kirk 2007: 1636). See also Berkson (1938), Cohen (1994), and Vicente and Torenvliet (2000), among others.

Accordingly, the NHST verdict on the reality of complete neutralisation as H_0 and of incomplete neutralisation as H_1 is simple: all neutralisation cases eligible for a phonetic investigation are a priori *incomplete*. Is it this capacity of NHST that led Dinnsen (1985) to suggest some twenty-five years ago that all putative neutralisation effects are in fact phonetically incomplete? I believe that it is rather the excitement that accompanied the relatively new ‘discovery’ of incomplete neutralisation, and the fact that there were (and still are) more studies reporting

⁷⁷ The expression ‘no systematic differences between A and B’ is just a variation on the theme and clever wording. It is statistical significance that is made to wholly and single-mindedly define what constitutes a systematic difference in our field. A difference that is statistically significant is considered systematic, whereas a difference that fails to achieve statistical significance is considered non-systematic.

⁷⁸ There are some researchers who protest that the claim that null hypotheses are always false only applies to point-null hypotheses of no-difference or no-correlation (Bakan 1966; Hodges & Lehmann 1954; Nickerson 2000; cf. Hagen 1997). For example, interval null hypotheses expressing a directional difference of an exact interval of non-zero values, while normally difficult to postulate, as defining such an interval requires a prior criterion of practical significance and many replicated data, can be true. Note that H_0 in neutralisation studies is a point-null zero-difference (or a difference that can assume any numerical value including zero) hypothesis whose statistical processing yields a test statistic with $p \geq .05$. Carver (1978: 381) assertively declares that “there is no way in practice that we can be absolutely sure the null hypothesis is true. If we could be sure, we would never test for statistical significance at all”.

incomplete neutralisation than studies reporting complete neutralisation. At the same time, I believe that a ‘publication bias’ cannot be possibly blamed for the prevalence of incomplete-neutralisation reports: as far as the phonetics of neutralisation is concerned, failing to reject H_0 has not been taken to mean that a study has failed and is thus unworthy of publication.

On the other side, many of the contributors to the empirical literature on neutralisation do not seem to be fully aware of the shortcomings of NHST. For example, the above description of the correlation between statistical significance and sample size naturally calls back to us a suggestion made by Warner et al (2004) and repeated in Warner et al (2006) that “such small effects as incomplete neutralisation require a large number of speakers and items in order to detect them (or rule them out) reliably” (Warner et al 2006: 290).

Let us now consider in detail how the Warner et al suggestion fares against a more realistic view of the limitations of NHST. Warner et al report that the statistical tests they ran on data from fifteen subjects in their 2004 study⁷⁹ yielded statistical significance for the voicing contrast, whereas those tests they ran on data from the same number of subjects in their 2006 study failed to yield statistical significance for the morphological distinction between verbs ending in /t/s and verbs ending in /t-t/s. How comparable are the findings of these studies? I will only consider this question from a statistical angle. We know from the discussion above that, with the kind of null hypotheses we postulate, any effect, however negligible, can be shown to be statistically significant if we recruit more subjects. The question to be asked is how many more subjects are enough.

To decide correctly, we need to calculate an index of effect size (I will say more about effect size further below). For the Warner et al studies, the first thing to note is that the mean differences in both of them are extremely minute. Specifically, a vocalic durational difference of 3.5ms in the 2004 study is found to be statistically significant, whereas a durational difference of 1.1ms between phonemically long vowels in verbs ending in /t/s and verbs ending in /t-t/s in the 2006 study is not.

⁷⁹ The Warner et al (2004) study also reports another statistically significant vocalic durational difference between hetero-graphic items with no phonemic distinctions. The difference is of a similar magnitude as in the voicing contrast (i.e., 3.4ms). For the sake of simplicity, the discussion above only refers to the voicing contrast in that study. However, the arguments given above apply to both distinctions.

However, it must be stressed that a small mean difference does not necessarily stand for a small effect size. A minute mean difference between two groups or two conditions with very little variability can make a large effect size. Mean differences *and* variability are both involved in the calculation of an effect size (see chapter four).

Now, since the two studies above do not give any measure of dispersion like SD, we are left to wonder if the lack of statistical significance in the 2006 study is due to the smaller mean difference, or if it is due to a more highly variable dataset in the 2006 study, or indeed to both. The post hoc power figures Warner et al give are not helpful at all. For example, the authors reported that the 2004 study “had power of .81 to detect the vowel duration difference (3.5ms) in that experiment [...while the 2006 study] should have a similar power to detect effects of comparable size” (2006: 291-292). But we do not know if 3.5ms and 1.1ms differences would make comparable effect-sizes. If these differences had comparable SDs, then we would need a larger-n study to ‘detect’ the 1.1ms difference at similar statistical power (see Nakagawa & Foster 2004 for arguments against using statistical-power analyses to interpret null results).

A conclusion one could more appropriately draw from the Warner et al studies would be to say that a sample of fifteen speakers was enough to secure statistical significance for the 3.5ms duration difference but not enough for the 1.1ms difference. The important point here is that having equal samples, by itself, is neither a sufficient nor necessary criterion of valid comparison when the relevant studies report different effect sizes. Having equal sample sizes in the Warner et al studies has neither established incomplete neutralisation for the 2004 study beyond doubt, nor has it reliably ruled it out in the 2006 study, contra the authors’ claim.

More importantly, it seems to me that accepting Warner et al’s logic creates a dilemma in defining just what a large-enough study is. For example, there are very small-n studies, sometimes even single-case studies, reporting incomplete neutralisation (e.g., Baroni & Vanelli 2000). These studies stand in sharp contrast to relatively larger-n studies that report complete neutralisation (e.g., Warner et al 2006). Perhaps a better claim might be that achieving statistical significance in a small-n study could be taken to imply statistical significance in a larger-n study,

whereas failing to achieve statistical significance in a large- n study would spell statistical non-significance in a smaller- n study (cf. Bakan 1966). Although this proposal merits further investigation, there may be cases where a small- n study finds statistical significance that disappears in a larger- n study. For example, an attenuated estimate of variability is more likely in a small- n study. Less variability makes statistical significance more likely.

The dilemma facing the Warner et al suggestion is actually resident in the whole NHST-dominated paradigm of the phonetics of neutralisation. We, as researchers subscribing to this paradigm, need to be specific about what should be a reasonable sample size. Should we follow a rule of thumb or calculate a priori power for the kind of tests we intend to run? Do we yet have the necessary statistics to perform a priori power, such as a measure of effect size from previous studies, for instance? Do we need to conduct pilot studies to get effect size figures if we cannot yet get them from the literature? Is there any point in going through all of this, just so that we have the right sample size for the effect to be detected at the level of power and the level of significance we have pre-selected? If, after all of this, no effect is detected, can we be confident that we have not missed it because it is not there?

Unfortunately, the question-begging in the preceding plea, though overlooked by many researchers dealing with the phonetics of neutralisation, does not help our case. Let us not ask how an effect is missed, but how it is detected. Here let me repeat Hubbard and Lindsay's (2008) conclusion that a test statistic with $p < .05$ provides no guarantee at all that there is a real effect. All that $p < .05$ says, as made clear above, is that there is a less than .05 probability of getting the sample statistic our study has yielded, or some more extreme ones, assuming that there really is no effect in the population.

Since, very crudely, mean differences, SD values, and sample size all contribute to the calculation of the commonly-reported t -value and F -ratio, a correlation between these attributes on the one hand, and the p value associated with these test statistics, on the other, is naturally expected. However, as pointed out above, the close correlation between statistical significance and sample size can also mean that a statistically significant difference might only be a *size effect* rather than an effect size. Effect sizes, in contrast, are not directly affected by sample size;

some effect size indexes can, in fact, be calculated independently of the sample size (e.g., Cohen 1988).

So far, I have discussed and illustrated three related problems with NHST. Firstly, statistical significance is easily misinterpreted. Secondly, the null hypothesis of no difference is almost always false on both logical and empirical grounds. Thirdly, statistical significance is highly sensitive to sample size.

These shortcomings of NHST should make us heed better long-standing warnings that statistical significance and practical significance can be different things (see e.g., Berkson 1938; Gold 1969; Kirk 1996; Tryon 2001). In fact, confusing statistical significance with practical significance has been yet another source of criticism levelled against NHST. Although some researchers lament the unfortunate choice of the word 'significance' whose non-technical use connotes importance (e.g., Meehl 1997; Schafer 1993), others insist that such a confusion is deliberate and has nothing to do with the inherent misinterpretability of statistical significance discussed at the beginning of the section. For example, commenting on significance testing, Bakan (1966: 423) protests that "a great deal of mischief has been associated with its use". Cohen (1994: 1001) spells that out:

All psychologists know that *statistically significant* does not mean plain-English significant, but if one reads the literature, one often discovers that a finding reported in the Results section studded with asterisks implicitly becomes in the Discussion section highly significant or very highly significant, important, big! [italics his]

It would be unfair, though, to think that Cohen's observation should be limited to psychologists. Pedhazur and Schmelkin (1991: 202) express irritation at what has become "common practice to drop the word 'statistical' and speak instead of 'significant differences,' 'significant correlations,' and the like". I myself have carried out a simple observational survey of most of the published studies on the phonetics of neutralisation and found that qualifying significance as statistical significance is the exception rather than the rule.

Some of the readily obvious signs of this unwanted confusion in the experimental literature on neutralisation and phonetic research in general is when the p value is the only statistic a study reports. This is what we find in some of the early studies of the phonetics of neutralisation (see e.g., Rudin 1980). Reporting the t -value or

the F-ratio is to be seen as an improvement (see e.g., Jassem & Richter 1989). But this is not enough. Disclosing more information about statistical significance by providing, for example, the degrees of freedom that the t-value and/or the F-ratio are assessed against is again a welcome gesture, but we should demand more (see e.g., Dinnsen & Charles-Luce 1984; Port & Crawford 1989; Slowiaczek & Dinnsen 1985).

Currently, the common practice is to give all the essential components of inferential statistics but little descriptive statistics. Inferential statistics mostly tells us about statistical significance, but very often, nothing else. In contrast, we can use descriptive statistics to form an impression of how big an effect is, how stable it is, and what direction it takes. This might inform our decision about the practical significance of the observed effect. Now it must be remembered that to calculate some standardised measure of this, such as Cohen's d , we can always use descriptive statistics.⁸⁰

The differential treatment of inferential statistics and descriptive statistics that we find in our discipline reflects what seems to be an institutionalised attitude whereby researchers attach priority to statistical significance and take it to be an objective and rigorous calibration of the more subjective practical significance. This attitude seems to be so deeply-rooted that even the critics of NHST can occasionally display. For example, Gold (1969: 46) contends that "statistical significance is only a necessary but not sufficient criterion of importance". Unpacking his statement, we must object to Gold's definition of importance: an important (practically significant) finding is, according to Gold, necessarily statistically significant, whereas a statistically significant finding is not necessarily practically significant. This is not always the case. Statistical significance is neither a necessary nor sufficient criterion of importance. It is rather practical significance that is both a necessary and sufficient criterion of importance.

It is ironic that in assessing what seems to be essentially subjective, equivocal, and vague practical significance, researchers have been and still are relying heavily on statistical significance in the mistaken belief that it is unequivocal, scientific, and

⁸⁰ To use the t-value or the F-ratio to calculate d , we need to make sure that these have been generated correctly. Sometimes, this is not an easy task, especially in Repeated-Measures designs (see Thalheimer & Cook 2002).

objective. The discussion throughout this section has shown that statistical significance can be subjective, equivocal, and vague. Moreover, among the reasons that Carver (1978: 393) lists to explain “the popularity” of NHST despite repeated criticism is that the “complicated mathematical procedures [underlying NHST] lend an air of scientific objectivity to conclusions”. Bakan (1966: 436) warns against this kind of unconditional esteem for mathematics and stresses that “[w]e must overcome the myth that if our treatment of our subject matter is mathematical it is therefore precise and valid. Mathematics can serve to obscure as well as reveal”.

With specific reference to phonetic research, one can hardly mistake the sustained fascination with mathematical modelling of phonetic data. Perhaps the undue obsession with statistical significance and the subordination of practical significance that we find in the experimental literature on neutralisation are but an expression of that fascination with mathematics.

COPYRIGHT MATERIAL

It seems to me that many researchers in our field have never had to take the distinction between statistical significance and practical significance very seriously, probably because they always seemed to have some explanation of their data. But let us not forget that the phonetic literature in general does record a few cases where an appeal has been made to this distinction. For example, Coleman (2003), confronted with statistically significant acoustic differences between 'lap' and 'Lapp' as produced by one subject, wondered whether statistical significance should necessarily imply linguistic (i.e., practical) significance.

Phonologically and phonetically speaking, what would be practically significant? Not trying to understate the enormity of this question, one can always give the following as a spontaneous answer. Phonologically and phonetically practical significance can be determined by theoretical considerations and by the knowledge we possess of the synchrony and diachrony of a language before we embark on experimentation. However, we should not forget that the theories we develop might be lacking in various respects, that our knowledge of the synchronic phonology of a language might be out of date, and that many parts of the diachronic picture of a language can sometimes be out of our reach.

Apparently, the question above should be approached from different angles. Yet, it seems to me that there will be times when a clash arises among the different answers we suggest. To illustrate, let us focus once again on the phonetics of neutralisation. Here, one might come across suggestions that production differences should be practically significant in their own right, irrespective of whether or not they are at all perceptible. At the same time, there are researchers who will insist on the perceptibility criterion for the definition of linguistically practical significance. A third category of researchers may be willing to disregard any production-perception correlation or lack of it as long as speakers' performance reflects what can be defined as lexical biases.

Perhaps what all of these three views share is that it is statistical significance that determines what effects are practically significant production differences, practically significant perception differences, and practically significant lexical biases. In other words, once our tests yield statistical significance, we rarely hesitate to announce practical significance, irrespective of whether we do

production data only, perception data only, both production and perception data, or lexical-decision data.

The question which urges us for an answer after this discussion of statistical significance and practical significance is this: how do we proceed with our empirical research questions? Actually, Morrison and Henkel (1970: 311) have volunteered an answer that assumes that statistical significance tests have no say in our research paradigm:

What we do without the tests, then, has always in some measure been done in behavioral science and needs only to be done more and better: the application of imagination, common sense, informed judgment, and the appropriate remaining research methods to achieve the scope, form, process, and purpose of scientific inference.

However, it seems to me that adopting the Morrison and Henkel approach can be just as radical as submitting wholly to NHST. Both approaches appear to me to be equally disastrous to the development and transmission of what needs to be cohesive science and knowledge. Perhaps, it is more desirable to allow our scientific inference to take its cue from both statistical reasoning that is based on properly interpreted statistics and the intuitive understanding of the relevant phenomena in the world.

For example, concerning the phonetics of neutralisation, one sensible way to do that is to think of our data as a variability composite; our task, then, is to decompose that seemingly amorphous whole into its component variability-sources, revealing at the same time their grouping structure, if applicable. We, thus, need to identify how much variation is due to our sample subjects; how much is attributable to the test items we employ; how much is explainable by the contrast variables that we specifically manipulate; and how much variation remains unexplained by the armoury of fixed and random factors we consider in the analysis.

The brief description of this statistical treatment is nothing more than adding efficiency to ordinary regression models. Unlike NHST procedures, model-fitting using regression equations seems to have evoked little criticism and is considered 'serious' statistics (Ableson 1997; see also Granaas 1998). Mixed or multi-level data modelling can paint a more detailed picture of the variability in the data and

its structure than can single-level models. In fact, multi-level modelling of linguistic data has been recently attempted (e.g., Baayen et al 2008; Quené & van den Bergh 2008, 2004). However, to do that appropriately, we will need to understand the underlying model, its requirements, and its limitations. Also, and perhaps more importantly, we need to know our data.

In this respect, it might be more rewarding to refine our understanding of phonetic variability than to fit mixed regression models to a set of phonetic data that has undergone experimental manipulation. Such manipulation can add to (or possibly obscure) the variability of these data.

I do not fit regression models to my data for inference here. Instead, I devote the rest of the thesis to sketching a new approach to variability. My contention is that an adequate understanding of variability and of our data should be attempted before seeking to get a ruling on how successful our model fitting has been.

Under this new approach, the empirical questions we will have in our study of the phonetics of neutralisation can find formal as well as intuitive answers in terms of variability decomposition. For example, we can learn more about how the various linguistic and non-linguistic influences impact on variability. I give more details in the next chapter.

To close, however, I note that, generally, nothing can settle the diverse empirical questions across the spectrum of scientific inquiry (including the phonetics of neutralisation) more satisfactorily than drawing conclusions based on replicated rather than individual studies. The key requirement here is for these replications to report some standardised measure of effect size and robust summary statistics (see the next section). These will facilitate meta-analyses. Only then are we justified in generalising beyond the sample subjects we have, without there being a need for us to take the potentially misleading path of NHST, where a randomly sampled set of data is formally processed to derive indirect evidence for or against a null hypothesis about the population from which the sample is supposed to have been drawn.

Now, for the majority of us language researchers, who sample at convenience rather than at random, and are comparatively tolerant toward the violation of a number of NHST formal assumptions, NHST-style generalisations from sample to

population may understandably inspire uneasiness. Well, this should make it all the more desirable for us to open up to other analytical techniques, including those assuming subjective and fuzzy approaches which allow intuitions to guide our scientific inferences. As a first step, we need to re-consider the type of summary statistics we use to quantify linguistic data. I turn to this next.

5.3 Quantifying the Phonetics of Neutralisation: Absolute and Relative Measures

5.3.1 *Measuring Central Tendency and Dispersion*

In many fields of research including ours, the most common measure of central tendency is the mean (\bar{X}), while the most common measure of variability is the standard deviation (SD), which can also be scaled to the mean in what is known as the Coefficient of Variance⁸¹ (C_V).⁸² C_V is said to be a useful measure of variability (Howell 2002), and one that is more consistent with people's tendency to think of variability relative to a measure of location (e.g., Lathrop 1967; Lovie & Lovie 1976). Most researchers assume that this measure of location is the mean (but see below for a different view).

There are, of course, other well-known measures like the median and the mode for locating central tendency, and the range, interquartile range (IQR), mean absolute deviation, and variance for estimating variability. For reasons of space, I limit the discussion to the most common measures listed in the paragraph above. But I will

⁸¹ I referred to this as the Relative Standard Deviation (RSD) in the preceding chapter. From now on, I will use the term Coefficient of Variance (C_V).

⁸² I gave the formulas for calculating the paired version of these variability measures in chapter four. Below I give their uncorrelated version, and for the convenience of the reader, I reproduce the equation for calculating the mean:

$$\text{Mean } (\bar{X}) = \frac{\sum X}{n} \quad \text{where } X = \text{an individual data point; } n = \text{number of data points}$$

$$\text{SD} = \sqrt{\frac{\sum_{k=1}^n (X_k - \bar{X})^2}{n-1}}$$

$$C_V = \frac{\text{SD}}{\bar{X}}$$

Note that the SD equation above is the one used for estimating the population parameter based on sample statistics. If SD is used solely to describe variability within a sampled dataset, the divisor in the equation should more appropriately be (n) rather than ($n-1$). However, statistical packages standardly use the equation with the correcting factor (i.e., $n-1$), regardless of what the experimenter's purposes are. The difference is usually very small, anyway (see Howitt & Cramer 2000).

discuss the mode and the range towards the end of this section. For more, see standard statistics books including Tukey (1977) and Howell (2002).

As its calculation makes use of all the data points in a set, the sample mean is said to be a sufficient estimator of the population mean (see e.g. Howell 2002). However, this also makes the mean very sensitive to outliers and extreme values, which can have a drastic effect on its robustness and resistance as a measure of location. Deriving the mean algebraically through summation and division ensures its admission into many equations that are necessary for estimation, which partly explains the wide use of the mean in inferential statistics (ibid). However, this property of the mean also reduces the chances of finding it as an actually occurring datum in a given dataset. In the overwhelming majority of cases, the mean remains an arithmetically derived value.

As with the arithmetic mean, the calculation of the SD uses all the data points within a set. This increases its sufficiency as an estimator of the variability of the underlying population, but, at the same time, it renders it vulnerable to the distorting effect of outliers and extreme values. Moreover, and just as with the mean above, the mathematical derivation of the SD is a point in its favour, statistically speaking. But being derived through subtraction, squaring, summation, division, and square-rooting makes the SD very non-transparent and unintuitive as a measure of variability. In fact, a number of researchers have mentioned the enormous difficulty that beginning students of statistics experience in trying to learn the concept of the SD as a measure of variability (see e.g., delMas & Liu 2004). I will return to the intuitiveness issue below.

As far as human cognition is concerned, there are suggestions in the literature undermining the cognitive plausibility of the arithmetic mean as a measure of average and the SD as a measure of variability. For example, Peterson and Beach (1967) observe that people tend not to take into account all the data they receive while drawing inferences. Peterson and Beach describe this behaviour, which is statistically sub-optimal, as conservatism. Kareev et al (2002) found conservatism in their subjects' estimation of variability. This indicates that the statistical advantage of using all the data points in the calculation of the mean and the SD does not find an obvious parallel in cognition.

Another example of the lack of correspondence between formal statistics and human cognition, involving again the mean and SD, concerns the fact that these measures, being arithmetically derived, are not naturally occurring values. This is highly relevant in the context of a conceptualisation of the human cognition as being sensitive to the rate at which events occur naturally in the world. Gigerenzer and colleagues (e.g., Gigerenzer 1993; Gigerenzer & Hoffrage 1995; Gigerenzer & Murray 1987) provide compelling arguments, based on both theoretical and experimental evidence, that the human cognition is designed to pick out frequency information, and that natural occurrences or underived frequencies (rather than derived probabilities, as currently thought) are all we need to make Bayesian inferences.

Gigerenzer and Hoffrage (1995) and Zacks and Hasher (2002) review numerous studies demonstrating the difficulty that physicians and statistics students experience when asked to estimate Bayesian posterior probability using data provided in a standard probability format. For example, a study by Eddy (1982, cited in Gigerenzer & Hoffrage 1995: 686) reveals that, of a sample of 100 physicians, 95 grossly over-estimated the 7.8% posterior probability $p(\text{cancer}|\text{positive})$ to be 70%–80%. Conversely, when test data are presented as frequencies, subjects have no difficulties arriving at the appropriate estimates. Gigerenzer and Hoffrage (1995: 686) argue that physicians' failure is not because "the human mind does not reason with Bayesian algorithms" but because probability format is the wrong format for cognitive tasks. The human mind is "tuned to frequency formats, which is the information format humans encountered long before the advent of probability theory" (ibid: 697). As Zacks and Hasher (2002: 21) put it, "people of all ages and under a very broad range of circumstances reliably and unintentionally encode information about the relative frequencies of events".

Importantly, for our case, the central tendency measure that is inherently associated with frequency is not the mean but rather the mode, which is the most frequently occurring value within a dataset. By definition, then, the mode actually exists as a naturally occurring datum. It is not arithmetically derived, as opposed to the mean (see e.g., Howell 2002; Tukey 1977). In other words, both comparatively high frequency and actual occurrence are essential components of the definition of

the mode. This should, in principle, make the mode more cognitively plausible than the mean as a measure of central tendency.

Furthermore, the mode, again as opposed to the mean, reflects another cognitive property. Human cognition is known to be geared to salient data (e.g., Carroll 2006). Saliency implies actual occurrence, and it can be envisaged as encompassing frequent occurrence as well. In other words, the two defining characteristics of the mode listed above further favour the mode over the mean as a measure of central tendency on grounds of cognitive plausibility.

Importantly, since the mode, again by definition, ignores infrequent values within a dataset, it is consistent with the conservative cognitive performance that Peterson and Beach (1967) describe (see above). This property, which once again adds to the cognitive plausibility of the mode, renders it lacking in statistical sufficiency and almost unusable in formal statistical inferencing. However, we have seen in §5.2.2 that formal statistical inferencing, at least as based on NHST, can be unreliable. At the same time, statistical inferences are built on probabilities. If we convert to a Gigerenzerian view of cognition, we will find those arguments for clinging to the mean while dismissing outright the mode very unconvincing.

Another related point to consider is the fact that the concept of average as utilised in statistical thinking under the name of central tendency can differ markedly from its non-technical denotation. Specifically, the notion of the average, to the statistically uninitiated, is closer to the mode rather than the mean. For example, Mokros and Russell (1995) have explored children's understanding of what the average is. Comparing the performance of fourth, sixth, and eighth graders, they found that fourth graders equated the average with the mode far more often than either sixth or eighth graders, who were, instead, more keen on applying the formal algorithm for computing the arithmetic mean or the median, which they learned in school. Interestingly, Mokros and Russell (1995) noted that those learners who were trying to derive the arithmetic mean were mostly unsuccessful in finding the mean, either because they fed the wrong information into the formula or used the correct information within the wrong algorithm. The authors sum it up for that type of pupil: "giving up what [...they know] about the world in order to apply a procedure that resulted in unreasonable (and unrepresentative) results" (p. 29).

In a histogram of a perfectly symmetrical unimodal dataset, the arithmetic mean coincides with the midpoint of the modal interval (i.e., the highest bar). But in practice, these measures almost always represent different values. When researchers subject their data, especially continuous data, to statistical treatment, they almost always pick the mean (but never the mode) to measure central tendency. With continuous data, it is almost always the case that there will actually be no true mode: every data point may occur only once. That is, no two data points may be exactly equal, given the large number of decimal places we can consider for comparison. Thus, there will most probably be no single value that will have occurred more frequently than others.

However, there are ways around this problem. For example, Bickel (2002) proposes an iterative formula for finding the mode in continuous data. Roughly, his algorithm, called the half-range mode,⁸³ consists in dividing the whole dataset into two intervals, then picking only the data within the highest bin (the modal interval). Next, the data within the modal interval are plotted into two intervals and a new modal interval is picked. This procedure is repeated for the remaining data until there remain exactly two data points. The mode will be the arithmetic average of these two points.

According to Bickel, this algorithm is more robust than other mode estimators in the literature. It is important to note that Bickel (2003, 2002) found the mode to be far more reliable than the mean or the median for asymmetric distributions, and, in fact, no less reliable than the mean and the median for symmetric distributions. Importantly, Hudson et al (2007: 1810), exploring F0 mean, median, and mode values for forensic purposes, conclude that the mode “gives a truer indication of central tendency”. Wittels et al (2002) draw a similar conclusion regarding F0 mode for the purposes of monitoring emotional stress (see also Loakes 2006; Rose 2002).

It is also important to realise that the algorithms proposed by Bickel return a single data point rather than an interval. Working with acoustic data, as I am in this thesis, I should prefer to pick an interval rather than a single point. This preference

⁸³ Actually, Bickel (2003, 2002) and Bickel and Frühwirth (2006) propose a few other algorithms for finding the mode. For reasons of space, I only sketch his algorithm for finding the half-range mode, which bears the greatest relevance to my arguments in this section.

is not only for practical reasons (although I admit reducing a dataset to a few intervals will increase its manageability). There are cognition-related reasons on offer. For example, Tversky and her colleagues conclude, on the basis of a number of empirical studies, that people tend to discretise continuous data (see e.g., Tversky 2005; Tversky et al 2008; Zacks & Tversky 2001; Zacks et al 2001). Tversky et al (2008: 437), summing up the case, write:

Why the mind discretizes is a question that has answers on many levels. On the neurological level, neurons fire or don't. On the cognitive level, the continuous input is so rich and complex that much of it must be, and is, ignored.

Continuous data are possibly processed in terms of a series of discrete points, each representing a stack of the points within its proximity. This stacking of data points can be pictorially depicted by plotting a set of continuous data as a frequency histogram.

Moreover, infants' processing of quantities suggests the presence of some binning activity. For example, Xu and Spelke (2000) show that 6-month-olds are able to discriminate a two-fold increase in quantity but not less than that. Their finding has been replicated in Brannon et al (2004), Xu (2003), and Xu et al (2003). In other words, quantities falling short of that ratio are indiscriminable, treated as being the same quantity—belonging to the same bin. It seems reasonable to assume that infants' processing of quantities is subject to a binning procedure that follows Weber's Law and has a certain threshold of discrimination. I present and discuss evidence for this kind of binning for phonetic data in the next chapter.

Note also that even Bickel's (2002) half-range algorithm sketched above, employs a procedure where intervals of progressively smaller widths are defined. It is interval width that needs to be treated with care here. It is important to realise that in dividing a set of data into intervals, the criterion of how wide an interval should be cannot be always meaningful for all types of datasets. Determining how wide intervals should be has actually been a much debated, yet still largely unresolved question (Scott 2009).

Since its publication in 1926, Sturges' rule⁸⁴ for class width, which forms the basis of the default setting of bin width in many statistical packages (Scott 1992: 48), has

⁸⁴ Sturges' rule is calculated as follows (in Scott 2009: 303): $K = 1 + \log_2(n)$

undergone a lot of modifications and optimisations to meet the new challenges that have been identified. Alternative formulas have also been put forth in statistics literature. Some of these rules employ some mathematical processing of the sample size (Sturges 1926), SD (Scott 1979), IQR (Freedman & Diaconis 1982) (all cited in Scott 1992). See Hirai (1989) and Scott (1992), (2009) for a review.

Defining interval width is certainly not a trivial exercise when the purpose of drawing a frequency histogram with a specific class width is not just for a better graphical illustration of the distributional properties of the data at hand. Finding the modal interval and defining the range of intervals within a dataset can be affected by how wide the data intervals are. Therefore, we need to find a criterion that we can apply consistently, but preferably, not on an ad hoc basis. Beyond certain statistical considerations (e.g., the underlying distribution of the data), choosing among the proposed formulas still has an element of arbitrariness. Of course, the consistent application of even an arbitrarily chosen criterion can reduce a great deal of the objectionability of its arbitrariness. However, a criterion that better suits our enterprise of understanding phonetic phenomena is one that we can aspire to justify in the particular context of phonetics. I describe and justify a phonetics-based criterion in the next chapter.

But before I conclude this section, I would like to return to the measure of variability that relates SD to the mean, i.e., the Coefficient of Variance (C_V). Derived as it is, C_V will automatically inherit any imprecision of the components that are fed into its equation. Furthermore, being a relative measure, C_V is actually tied to the numerical value of its components, irrespective of the range of possible values which those components may assume.

More concretely, the proportion of a hypothetical SD of, say, 10 to a mean of 50 within a dataset with a range of 30 is far larger than the proportion of the same SD to a larger mean of, say, 500 within a dataset with an exactly equal range—30. In the first case, C_V is .2, with SD forming 20% of the mean. In the second case, C_V is .02, with SD representing only 2% of the mean. That is, according to C_V , the first dataset is much more variable than the second dataset. By contrast, SD and range figures alike indicate that the two datasets are equally variable. One point of difference in this comparison is that C_V is not independent of the numerical value of the mean. For this reason, C_V can sometimes give a wrong estimate of

variability. I illustrate this point using fictitious datasets with the same distributional shape as in Figure 5-1.

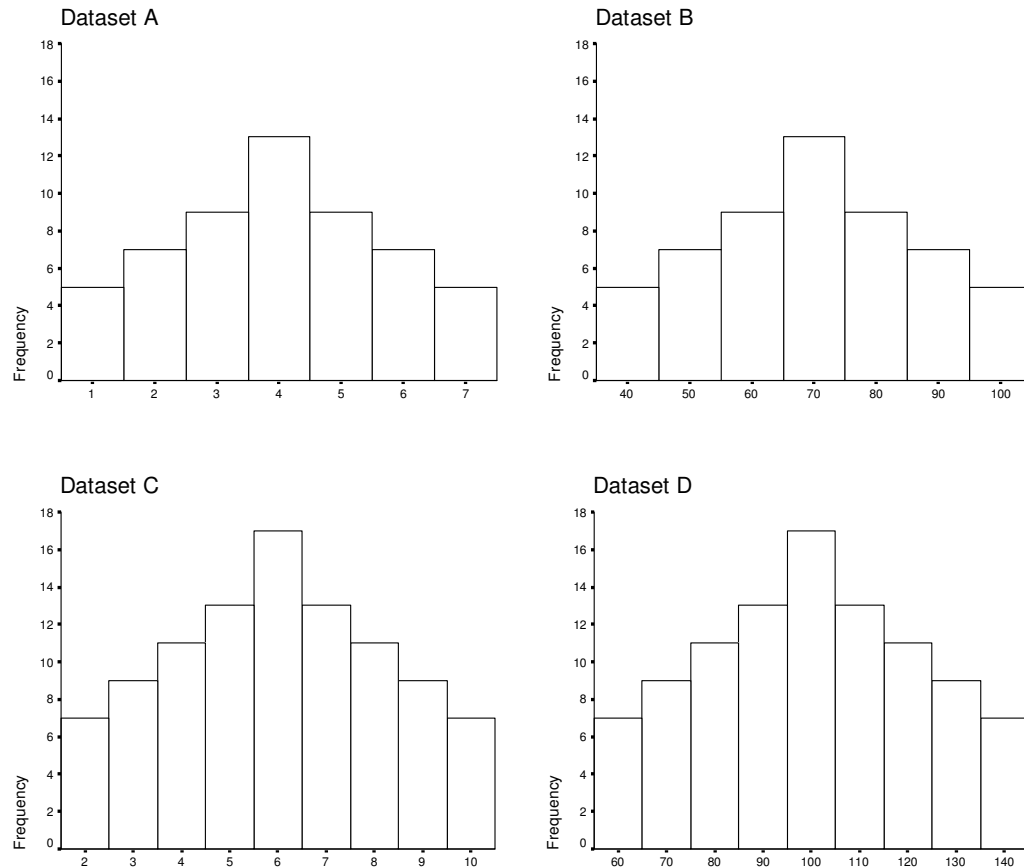


Figure 5-1: Histograms of four fictitious datasets with the same distributional shape

	Dataset A	Dataset B	Dataset C	Dataset D
Mean	4	70	6	100
SD	1.74	17.4	2.28	22.8
C_v	.43	.25	.38	.23
Range	6	60	8	80

Table 5-4: Mean, SD, C_v , and range values of the four fictitious datasets in Figure 5-1

As we can see from Table 5-4, all the values (mean, SD, C_v , and range) are different among the four datasets. How variable these datasets are seems to depend

partially on the measure we use. The above measures seem to paint a different picture of variability for these datasets, as shown in Figure 5-2 below. For example, using the SD or the range⁸⁵ as our measure of variability, we conclude that no two sets have equal variability, with dataset (D) being the most variable while dataset (A) being the least variable, and with dataset (B) being more variable than dataset (C). According to C_v , (D) is the least variable, while (A) shows the most variation. Of the middle histograms, (C) is more variable than (B). See Figure 5-2.

	SD	C_v	Range
Most variable	D	A	D
	B	C	B
	C	B	C
Least variable	A	D	A

Figure 5-2: Variability in the four datasets in Table 5-4 as measured using SD, C_v , and range values

The problem with these measures is their reliance on the numerical values of the components of their formulaic definitions. For example, the range is sensitive to the numerical values of the smallest and largest data points within a set. A dataset ranging in units is very likely to have a smaller range than a dataset ranging in tens, which in turn is expected to have a smaller range than a dataset ranging in hundreds, and so on.⁸⁶ Similarly, the SD is sensitive to the numerical value of the distance between the mean and each of the data points, especially extreme ones in a set. The larger the numerical value of this distance, the larger the SD gets (Howell 2002). Finally, the datasets illustrate very vividly how C_v is sensitive to the numerical value of the mean. It is clear that a relative measure of variability that is less dependent on the numerical values of the absolute measures of central tendency and variability should be more desirable for quantifying variability. Below I suggest a measure that has the potential of being just that.

⁸⁵ Note that if we express the range, not in terms of real data values, but in terms of bar numbers, we will reach the conclusion that datasets (A) and (B) have equal variability, that datasets (C) and (D) are equally variable, and that the first two sets vary less than the second two sets. Interestingly, if we choose to relate the range values to their corresponding means, we will come to exactly the same conclusion as drawn on the basis of C_v .

⁸⁶ Exceptions to these are also likely, but very much less so. Of course, mathematical normalisation is one way to circumvent that problem. But normalisation or transformation does not seem to be a pre-requisite for the majority of the statistical models we run. On the contrary, many statisticians recommend that transformation be kept to a minimum, as it tends to complicate the subsequent interpretation of the data (see e.g., Osborne 2002).

5.3.2 Towards an Intuitive Notion of Variability Fields

5.3.2.1 Preliminaries

According to Pingel (1993: 71), “the concept of variability is not defined precisely enough to lead to a single measure of variability”. The variability measures discussed so far are but a few of the proposed variability measures. As we have seen above, different measures of variability suggest different answers to the question of how variable a dataset is. Occasionally, a discrepancy arises between what statistics says and what intuitions say regarding the size and definition of variability. For example, all the datasets in Figure 5-1 above intuitively seem to be of equal variability—a conclusion that is supported by none of the measures considered above (i.e., SD, C_v , and the range). Consider now a situation where the reverse scenario obtains. This is when a number of datasets intuitively appear to be different in variability, yet statistical measures of variability insist on seeing no differences among them. In Figure 5-3, I illustrate this curious case, using histograms “cleverly constructed” by Nitko (1983) and reported in Pingel (1993: 70-71).

COPYRIGHT MATERIAL

Figure 5-3: Four datasets with the same mean and SD values (adapted from Pingel 1993: 71)

COPYRIGHT MATERIAL

Table 5-5: Mean, SD, Cv, and range values of the four datasets Figure 5-3

All four sets in Figure 5-3, which have different distributional shapes, have, nonetheless, the same mean, the same SD, and of course, the same C_v (see Table 5-5).⁸⁷ By these measures, the four sets have exactly the same variability. However, this does not seem to agree with the intuitions of many people—certainly not with my intuitions nor with those of the sample I consulted. Intuitively, dataset (D)

⁸⁷ I discuss the range further below.

seems to be the least variable, as all of its data points except one stack into only two bins. Dataset (C) is the next least variable, as half of its data points fall into one bin while the remaining half scatter into four different bins. Datasets (A) and (B), which have the same amount of variation, are the most variable. In each of them, data points forming 67% spread out into three bins while the remaining 33% fall into a fourth bin.

Note that the conclusions based on SD and C_v do not agree with these intuitions. According to these measures, all four datasets are equally variable. These measures express, respectively, how widely data points spread out from the mean value, and what proportion that deviation forms relative to the mean. That is, they measure the degree of density around the arithmetic mean. This does not reflect the intuitive understanding of the notion of variability that people with little or no statistical training may have. I have indicated in §5.3.1 above that neither the arithmetic mean nor the SD is intuitively available to statistics-naïve subjects. This will, by extension, disqualify C_v as an intuitive measure of variability.

The intuition-based description of the different datasets in Figure 5-1 and Figure 5-3 above suggests that there are basically two elements whose variation appears to contribute to our intuitive definition of variability. These are the frequency of the modal interval and the range of bins in those histograms. The higher the frequency, the smaller the variability will be, whereas the larger the range, the larger the variability will be. This agrees with an observation made by Lann and Falk (2003) regarding students' approach to variability. These researchers report that the range scored the highest in all the conditions of their questionnaire. The other measures of variability they studied were SD, mean Absolute Deviation, and IQR. These measures were used very inconsistently by the subjects and scored only about half of the score of the range. Moreover, the verbal explanations that subjects had to provide at the end of the questionnaire about what exactly they considered while assessing the variability of the datasets actually contained an explicit reference to the intuitive correlation between the presence of repeating values and decreased variability.

It is clear from the findings emerging from Lann and Falk's (2003) study that students' intuitive conception of variability has an element of density and/or a component of spread, but that neither density nor spread is around the arithmetic

mean in particular. Density here is the frequency of the average in the intuitive sense—i.e., the modal interval; spread here can be the range of the intervals comprising the dataset. The variability index I propose next makes use of these measures.

5.3.2.2 The Variability Field Index (VFI)

VFI relates the frequency of the modal interval to the number (i.e., range) of intervals that a dataset falls into. Note that scaling the range of intervals to modal frequency frees this index from the numerical values of the modal interval as a measure of central tendency and from the numerical values of the range as a measure of variability. This is one point of difference between this index and other common measures discussed above.

Importantly, being expressed as a ratio, VFI captures the cognitive relativity that characterises our intuitive judgements of variability. As pointed out above, researchers have long noted people's tendency to assess variability relative to some measure of location (Lathrop 1967). Interestingly, Lovie and Lovie (1976) report that statistics-naïve subjects have some success in repressing this relativity following explicit instructions but only when variation is low. With a larger amount of variation, however, subjects just cannot override this tendency.

To the extent that VFI is just an approximation of the intuitive (not the statistical) notion of variability, we need to understand the nature of the contribution of its components to our intuitive conception of variability. On the one hand, the frequency of the modal interval seems to be in inverse proportion to the perceived amount of variability. As pointed out above, the frequency of the mode acts as a density effect. The higher the frequency of the modal interval, the more densely populated it is, and the less variability a dataset will appear to show. To model this, I take the frequency of the modal interval to represent the force of gravitation in a variability field. The higher the frequency, the greater the gravitation a variability field will display. Importantly, a greater amount of gravitation creates an illusion of a smaller variability field. This agrees with the claim that people think of the world as far less variable than it really is (Kareev et al 2002; Peterson & Beach 1967). That VFI is capable of formally capturing this illusion is yet another point in its favour.

On the other hand, the range of intervals defines the notion of spread. It is positively correlated with the perceived amount of variability within a given dataset. The greater the range, the greater the variability a dataset will appear to have. That is, the range of intervals and the frequency of the modal interval pull our perception of variability in opposite directions. Consider the two line graphs in Figure 5-4. They both have the same spread but differ with respect to gravitation, with (a) having far greater gravitation than (b). At a quick look, the dotted line in (b), which represents spread, appears as though it is longer than the dotted line in (a). This is a visual illusion partly caused by our assessing of spread relative to gravitation.

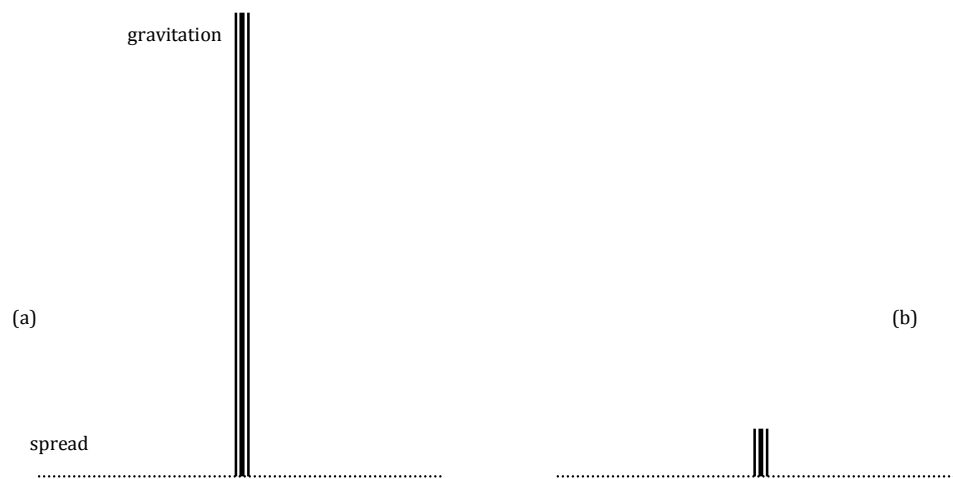


Figure 5-4: Line graphs with the same spread but different gravitation bars.

To appreciate how an increase in gravitation creates an illusion of a small field, we need to think of the gravitation bar in the chart above not as an area, but as an amount of force: the more data there are within the modal interval, the stronger the gravitation is, and hence the smaller the field will appear. To visualise this, consider the pyramid chart in Figure 5-5, where shading indicates the force of gravitation. The thick vertical line in the middle of the pyramid stands for the modal interval while the numbered arrows represent occasions, arbitrarily selected, for assessing variability at different modal frequencies. These schematic occasions are only meant for illustration and ease of reference. For example, on occasion (0), where the modal interval has not yet formed, no intuitive assessment

of variability may be possible. This is indicated by a solid arrow superimposed on the thick line representing spread. For the rest of occasions, an intuitive assessment of variability is possible. This is indicated by dashed arrows. The graph illustrates the inverse relationship between the frequency of the modal interval and perceived variability: as spread is kept constant in this graph, we can see that the greater the frequency of the modal interval, the smaller the perceived variability is. So perceived variability at occasion (1) is far greater than that at occasion (5), where the frequency of the modal interval is very high.

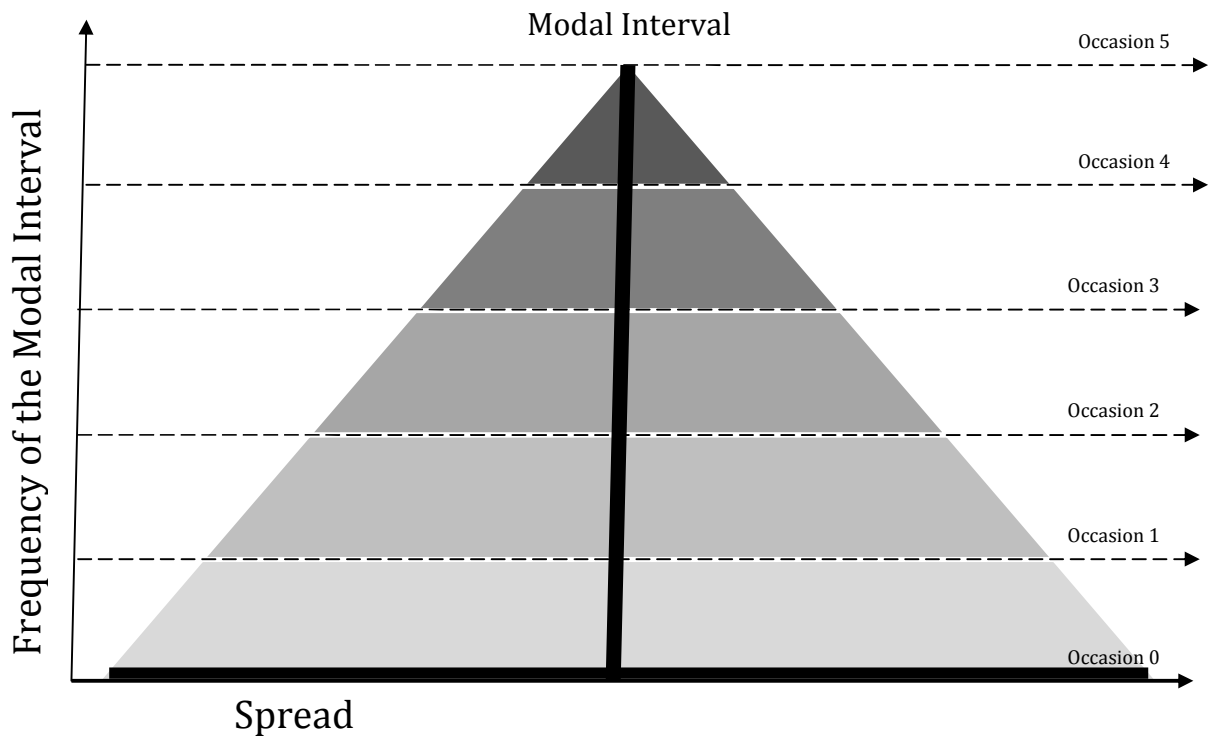


Figure 5-5: A pyramid graph illustrating the inverse relationship between frequency of the modal interval (shown as a vertical line in the middle of the pyramid) and perceived variability (indicated by shaded bands of different sizes); shading indicates the force of gravitation.

Thus far, I have described and illustrated the relationship between gravitation and spread, and their contribution to perceived variability. It is clear from the discussion that a formal measure of variability that is both intuitive and cognitively plausible must somehow relate gravitation to spread. Expressing that measure as the ratio of spread to gravitation will capture the shrinking effect that an increase in gravitation causes our perception of variability to have. The greater the

gravitation, the smaller the ratio will be. At a very crude level of conceptualisation, VFI might be reminiscent of the Coefficient of Variance (C_v) discussed earlier. Recall that C_v relates SD (spread) to the mean (central tendency). VFI relates the range of intervals (spread), not to the modal interval (central tendency), but rather to the frequency of the modal interval. That is, unlike C_v , VFI is not tied to the numerical value of the location measure or to that of dispersion. Frequency and bin counts can be independent of the numerical value of the measures of central tendency and variability, respectively. More formally,⁸⁸ the function to derive VFI is given in (49).

$$(49) \quad \text{VFI} = \frac{(Ni-1)}{(Fm-(Nm)^2)} = \frac{S}{Gf}$$

S stands for spread⁸⁹, Gf stands for gravitation force. Spread is defined as the number of intervals (Ni) minus 1, while gravitation force is the difference between the frequency of the modal interval (Fm) and the number of modal intervals (Nm) squared (i.e., multiplied by itself). The latter is a correcting factor. Since real data are not always unimodal, we need to correct for modality. This correction penalises the number of modal intervals such that the cost of modality increases rapidly as the number of modal intervals increases. Consider the hypothetical datasets in Figure 5-6.

⁸⁸ Thanks to Dr Jasmina Panovska-Griffiths for useful advice on the mathematics in the thesis.

⁸⁹ Defining the range of intervals as the difference between the number of intervals and 1 is not really different from the statistical definition of the range as the difference between the maximum and the minimum data points. In calculating VFI, we are not dealing with points but with the number of existing intervals. Consequently, the minimum of the number of existing intervals is invariably 1, whereas the maximum depends on the dataset under analysis.

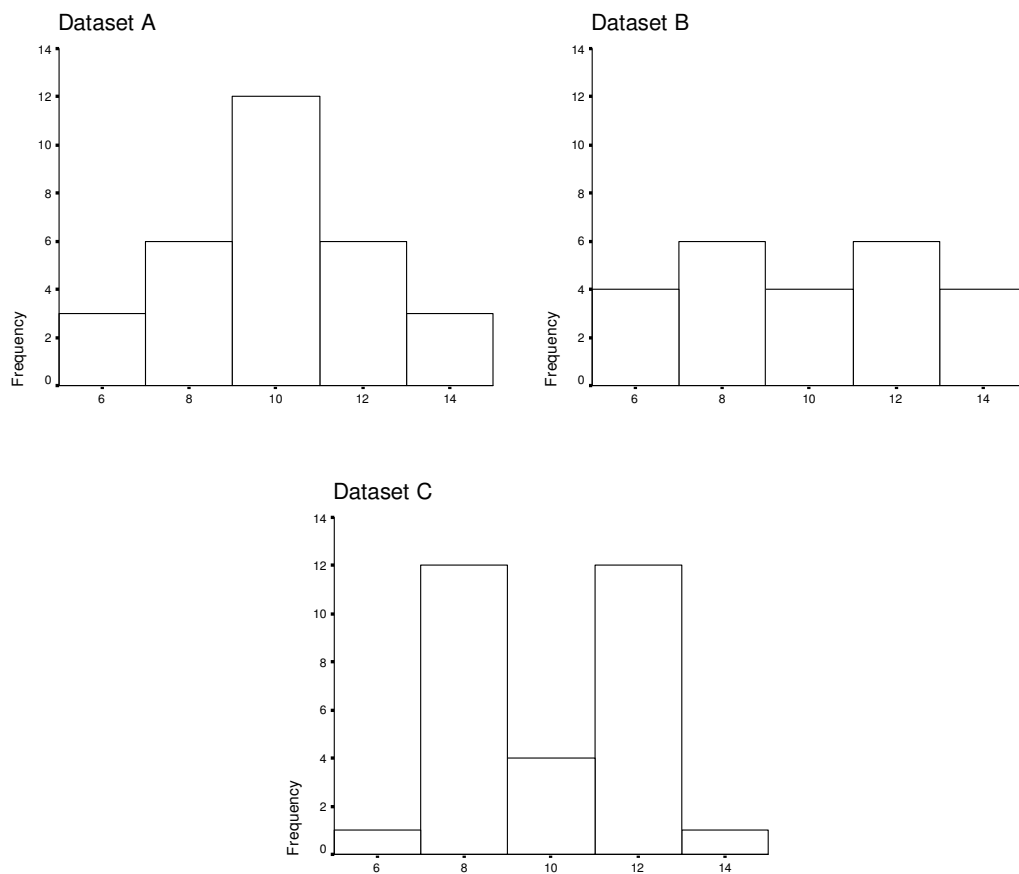


Figure 5-6: Histograms of three hypothetical datasets having the same number of intervals but differing with respect to their modal intervals

		Dataset A	Dataset B	Dataset C
VFI	Frequency of Modal Interval	12	6+6	12+12
	Number of Intervals	5	5	5
	Uncorrected	.3	.3	.16
	Corrected	.36	.5	.2

Table 5-6: VFI values of the three hypothetical datasets in Figure 5-6 with and without correcting for bimodality

It is clear from the table above that correcting for modality helps bring the values of VFI more in harmony with our intuitions.⁹⁰ Datasets A and B can't be equal in variability. In dataset A, twelve data points all stack up in one modal interval. By contrast, in dataset B, the same number of data points spread over two modal intervals. Adjusted for modality, VFI now declares that dataset B is more variable than dataset A and that dataset C is the least variable.

Next, let us compare VFI with the widely used measures of SD and C_v , using the histograms in Figure 5-1 and Figure 5-3. For ease of comparison, I have reproduced the relevant tables and added a row for VFI (see Table 5-7 and Table 5-8)

	Dataset A	Dataset B	Dataset C	Dataset D
Mean	4	70	6	100
SD	1.74	17.4	2.28	22.8
C_v	.43	.25	.38	.23
Range	6	60	8	80
VFI	.5	.5	.5	.5

Table 5-7: Variability measures applied to the four fictitious datasets histogrammed in Figure 5-1

	Dataset A	Dataset B	Dataset C	Dataset D
Mean	4	4	4	4
SD	1	1	1	1
C_v	.25	.25	.25	.25
Range	3	3	4	2
VFI	.75	.75	.67	.25

Table 5-8: Variability measures applied to the four fictitious datasets histogrammed in Figure 5-3

As we can see from Table 5-7, all measures of variability except VFI support a conclusion that all four datasets in Figure 5-1 have different variability. I expressed

⁹⁰ Note that ignoring bimodality and considering the frequency of only one modal interval in the histograms in Figure 5-6 still results in an estimate that is at odds with our intuitions. Specifically, datasets A and C will then have the same variability, and will be equally less variable than B.

the intuition earlier that these datasets are equally variable. VFI formally captures that intuition.

Next consider the datasets in Figure 5-3. As suggested earlier, treating these four datasets as being equal in variability is counter-intuitive. By criteria of SD and C_v , these four datasets have the same variation. This is not the conclusion we get by consulting the range or VFI. However, the picture VFI paints is more congruent with our intuitions than that of the range. The difference lies in whether or not datasets A and B are more variable than dataset C. According to the range, C is more variable, because the range only takes into consideration the most extreme values, and nothing else besides them. However, VFI, which takes into consideration both the range of intervals and, importantly, the frequency of the modal interval, declares that datasets A and B are more variable than C. This agrees with the intuitions of the people I interviewed, and indeed with my own.

5.4 Conclusion

In this chapter, I presented a qualitative and quantitative description of the phonetics of neutralisation. On the qualitative side, I examined statistical significance, the single most important criterion in characterising the phonetics of neutralisation. I have shown how a statistically significant difference can be a size effect rather than an effect size. I have also shown that statistical significance can easily be misinterpreted, and that the null hypothesis of no difference is almost always false on both logical and empirical grounds.

On the quantitative side, I evaluated the parametric measures of central tendency and dispersion. I have shown that these measures are unintuitive and so closely tied to the numerical values of the measurement scale that a robust estimation of the underlying central location and variability can sometimes be undermined. I have also proposed an alternative that is both more intuitive and cognitively plausible. Moreover, I have suggested that phonetic data are more appropriately examined as intervals rather than as single points. In the next chapter, I sketch a general model that builds on these insights.

6 A Sketch of the Variability-Field Approach

6.1 Introduction

One of the notoriously difficult challenges facing researchers who are concerned with modelling speech perception, production, and processing is how to deal with variability in the sound wave (see Connine & Pinnow 2006; Elman 1992; Klatt 1992; Labov 1986; Luce & McLennan 2005; McMurray et al 2002; Nguyen et al 2009; Ohala & Feder 1994; Pisoni 1997; Sommers et al 1994, among many others). Approaches to the question of phonetic variability vary widely. At one end of the spectrum, the traditional theory of the phoneme looks for structurally-governed macro-events in the variable speech stream, disregarding as cognitively irrelevant any fine phonetic detail. Commenting on this approach, Tatham (1976: 48) writes:

That the phoneme covered a range of variations was never very difficult to agree upon. Precisely defining the range to be covered, however, was always guaranteed to precipitate considerable argument.

Bearing witness to Tatham's claim is the assortment of labels we have accumulated over the years to classify sounds into phonemes, quasi-phonemes (e.g., Scobbie et al 1999; Scobbie & Stuart-Smith 2008), intrinsic and extrinsic allophones (e.g., Ladefoged 1967; Wang & Fillmore 1961), quasi-allophones (e.g., Rose & King 2007), and deep allophones (Moulton 2003). This classification relies on the notion of contrastiveness, which has traditionally been established categorically but is now being attempted probabilistically (see Hall 2009 for more on this).

At the other end lie exemplar-based approaches (e.g., Goldinger 1997; Johnson 1997; Lachs et al 2003; Nosofsky 1986; Palmeri et al 1993; Pierrehumbert 2001, 2002; Pisoni 1997), which look for micro-events in the variable speech signal. Fine phonetic detail has never been more important both theoretically and practically,

whereas phonemic labelling has become secondary and at best an emergent, data-driven activity.

In between these two extremes are approaches focusing on the organicity (e.g., Elman & McClelland 1986; Klatt 1992), functionality (e.g., Johnson 2001), and dynamicity of variability (e.g., Nguyen et al 2009; Nolan et al 2006; Tuller et al 2008, 1994). Importantly, these approaches have the capability of addressing both the macro and micro dimensions of the issue.

Interestingly, however, the view shared by all of these approaches, and by many researchers in our field, is that even though a great deal of phonetic variability is not random, it remains unmanageably large and overwhelmingly pervasive. Klatt (1992: 218), for one, attributes the disappointing performance of contemporary models of perception and recognition to the fact that “[w]e simply do not know how to deal with the seemingly unreasonable variability” in speech (see also Luce & McLennan 2005). Although Klatt is mostly concerned with engineering implementation, his confession actually incriminates the inadequacy of our procedural, empirical, and conceptual perspectives on phonetic variability. In this thesis, I shall be mostly concerned with sharpening the conceptual side of our perspective on variability. However, for scope and time limitations, the approach I adopt is mostly illustrative.

It seems to me that a first necessary step toward addressing adequately the question of phonetic variability is to recognise that variability is the essence of phonetic data rather than some isolatable addition (cf. Johnson 2001). At the same time, variability can be composite as well as primary. The difference here is one of analysability, with composite variations being analysable into portions attributable to traceable causes. I detail this view in the next sections.

But I observe here that there seem to be a lot of confusion and misperception surrounding phonetic variability. For example, in dealing with variability, most researchers start with the premise that phonetic variability is considerable and ubiquitous, equating, at the same time, lawful variability with non-random variability. In this thesis, I suggest a very different view.

Specifically, I start by re-charting the territories of lawfulness and randomness as applied to variability in light of the claim that phonetic variability is structured,

constrained, and not at all large, contra current beliefs. According to this new approach, lawfulness and non-randomness are no longer interchangeable. The main difference between what is lawful and what is non-random is one of acceptability versus explicability. The former refers to what is acceptable to a language community, whereas the latter is decided by language experts. Obviously, what is acceptable and what is explicable may intersect but may not always completely overlap. Thus, variations falling within the bounds of acceptability, regardless of their explicability, are lawful. In contrast, variations caused by known sources are explainable, thus, non-random. As is clear, to get explainability judgments, we may need expert advice from linguistically non-naïve subjects. In contrast, we can get acceptability judgments even from naïve native users of the language under investigation. It is important not to confuse explicability with acceptability when we deal with phonetic variability. All forms of acceptable variability are eligible for inclusion in our domain of investigation of phonetic variability. That being the case, we should be prepared to welcome variations that we cannot explain, just as we welcome variations that we can explain. Explainable variability is, simply put, a research finding, if positive. We should not let it prejudice us against negative findings like unexplainable variability.

Next, I present a skeletal description of the Variability-Field approach, highlighting its underlying conceptual philosophy. I first explain how phonetic variation is both constrained and structured, supporting my claims with real and hypothetical data. I then show how the Variability Field Model (henceforth VFM) introduces a new perspective on the processing and representation of variability. I focus specifically on what is commonly known in the literature as allophonic and indexical variations (e.g., Abercrombie 1967; Ladefoged 1993; Luce & McLennan 2005; Nielsen 2008; Nolan 1983; Ogasawara 2007; Pisoni 1997). It is these sources of variability that, according to a very recent appraisal of the current models of speech perception and recognition, pose a real challenge to current and future endeavours to understanding and modelling speech activity. They are obviously in need of addressing (Luce & McLennan 2005). I discuss a VFM schema whereby allophonic variations form context-bound phone-fields, whereas indexical variations provide some form of background against which a phone-field is accessed.

The chapter proceeds as follows. In §6.2, I briefly describe the confusion that characterises our approach to phonetic variability. In §6.3, I present a sketch of VFM, summarising in §6.3.1 its main philosophy and describing the variability effects that are modelled in the chapter. In §6.3.2, I show how phonetic variability is both constrained and structured. In §6.3.3 and §6.3.4, I detail the components of VFM and describe the inter-relations among them. Finally, I sum up the main claims and conclude in §6.4.

6.2 Dealing with Phonetic Variability: The Confusion

Despite the rapidly growing recognition of the theoretical and practical importance of variability in phonetic research (see e.g., Lachs et al 2003; Luce & McLennan 2005; McMurray et al 2002; Nygaard 2005; Perkell & Klatt 1986; Singh 2008), it is still unclear as to what constitutes phonetic variability. Blache and Meunier (2004: 2) sum it up:

For a long time, variability was regarded as noise making obstacle to the identification of the sounds of language. This is a complex question because there is confusion about the different phenomena which one calls variability. Indeed, it is common to call “variation” phenomena as distinct as coarticulation, speech style or random variation. Confusion is due to the fact that one is unaware of what should be a realization with no variation. Thus, is coarticulation a special type of variation compared to prototypes of isolated phonemes? Or does it form an integral part of the prototypes of speech production?

There may be several causes for phonetic variability and for this confusion. Yet it seems to me that we will go a long way towards resolving this confusion if we come to realise that phonetic variability is the essence of phonetic data rather than some isolatable addition. In this sense, we should no longer ask if an effect is to be treated as variation or as an integral part of the speech stretch in question, implying that there is an inherent incompatibility between being an instance of alienable variation and being an inalienable component of speech. There should be no incompatibility or distinction: phonetic data are essentially a blend of variability sets, most exhibiting multiple sources of variation at play. Such multi-source variation is responsible for the composite, non-uniform character of

phonetic variability, which we should seek to make provision for when we attempt to estimate how variable phonetic data are. Our task should be to decompose these multi-source variability sets until we reach a stage where variability sets are not further reducible to smaller component sets traceable to known sources. Such will be primary variability. Here, there will be no disparity, no diversity. This is variability in its unadulterated form.

On this view, depending on the kind of questions we ask, we may want to look at variability in its totality or in its composite or primary forms. If our purpose is to form an overall, gross picture of variability, many might think that we probably need not decompose variability. But the picture will certainly be far less blurry with the inclusion of the details we get by studying variability in its primary form. Part of the confusion that marks our treatment of variability stems from our being unspecific about what variability form we are studying.

Another factor contributing to this confusion is the prior assumption, shared by many researchers, that exaggerates the magnitude and ubiquity of phonetic variability (see e.g., Connine & Pinnow 2006; Lavoie 2001; Nygaard 2005; Peterson & Barney 1952). Indeed, Sheila Blumstein (1986: 199) spoke for many when she wrote that “there is no question that there is a tremendous amount of acoustic/articulatory/phonetic variability in speech. This is not at issue”. But here I would like to take issue with that. Specifically, I claim that the perception that phonetic variability is rampant and unrestrained is an illusion created by confusing the different forms of variability described above and by certain conceptual assumptions and methodological practices that usually define our approach to variability.

One such inappropriate practice concerns the way we measure variability, which I have detailed in the previous chapter. Briefly, my main reservations are these. Phonetic variability is statistically measured relative to the arithmetic mean, which is usually an imaginary point constructed through arithmetic summation and division using all the data points within a given set of data. Arithmetically derived measures of variability such as the standard deviation (SD) and the coefficient of variance (C_v) are not cognitively plausible for at least two reasons. Firstly, using all data points within a set to calculate variability does not reflect cognitive conservatism, where only a subset of a given dataset is considered. This latter

mode of cognitive calculation, unlike SD and C_v , underestimates objective variability (e.g., Kareev et al 2002; Peterson & Beach 1967). Secondly, being arithmetically derived, these variability measures are almost never naturally-occurring values. As such, they may not be readily available to human cognition, which is said to be geared to events that occur naturally in the world (e.g., Gigerenzer & Hoffrage 1995; Zacks & Hasher 2002).

Another inappropriate practice involves the type of pre-analysis treatment that phonetic data undergo. Specifically, datasets are commonly treated as being composed of individual data points, instead of intervals. This practice seems to unduly and massively magnify the ubiquity and size of phonetic variability. Many researchers would agree that a great deal of phonetic variation falls below the threshold of perceptibility. Yet many would still include in their calculation of variability perceptible and imperceptible differences alike. More concretely, suppose we have for analysis a dataset of a hundred physically different data points, making up only ten perceptibly different tokens. There will be a size difference between the range of possible values measuring physical reality and the range of possible values measuring perceptibility bounds. See the next section for more on this.

A third factor leading to the creation of this confusion, but which has received little attention in the literature, is equating implicitly or explicitly lawfulness with non-randomness in our characterisation of phonetic variability.⁹¹ For reasons of space, I mention here only two views. The first view bestows lawfulness on variability that is predictable (see e.g., Elman & McClelland 1986). This is variability due to physiological or phonetic reasons—variability that is mechanical, to borrow a term from Johnson (2001). The second view sees as lawful any variations that are meaningful in the sense that they are linguistically relevant, be it syntactically, pragmatically, lexically, or socio-linguistically, etc. (see e.g., Johnson 2001; cf. Gafos 2006; McMurray et al 2005).

The mechanical-variability view legalises phonetic variation on organicity grounds, whereas the meaningful-variability view legalises phonetic variation on

⁹¹ It is important to note that approaches focusing on the dynamic nature of phonetic variability (e.g., Nguyen et al 2009; Nolan et al 2006; Tuller et al 2008, 1994) do not suffer from this shortcoming. It is for this reason that I do not discuss them here.

functional grounds.⁹² Note that the two views share an obvious appeal to explicability in defining what to consider and what not to consider as lawful variability. According to both views, variations that defy explanation remain outlawed. These variations include inter-trial variations and a portion of intra- and inter-speaker variations. The commonalities between the two views are summarised in Figure 6-1.

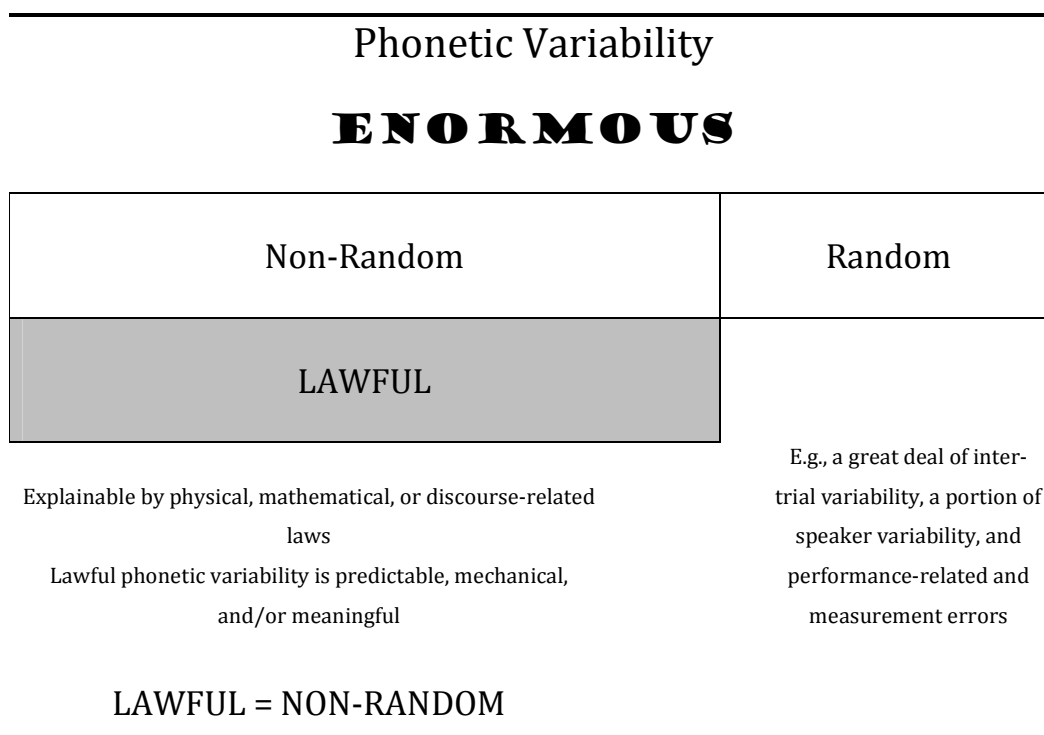


Figure 6-1: Variability as viewed in the literature: Lawful = Non-random

It appears to me that equating lawfulness with non-randomness can encourage disinterest in random variation, thus prejudicing our attempts to understand phonetic variability. If random variations are to be excluded from our domain of inquiry, they should be excluded on the basis of an adequate investigation rather than on a preconception. Until that happens, exploring random variations should remain a legitimate research topic. Meanwhile, it may be instructive to familiarise ourselves with the stance that other fields of inquiry take on random variability.

Let us focus specifically on brain research and complex systems. There, we come across demonstrations strongly supporting the claim that some amount of random

⁹² Note here that organicity and functionality are not necessarily mutually antagonistic. The exact relation between the two should be interesting but is actually beyond the scope of this thesis.

variability is not only beneficial but actually necessary for the optimisation of neural systems. Random variability has been shown to improve the responsiveness of those systems. For example, Levin and Miller (1996) demonstrate that sensory neurons perform better in the presence of noise. Specifically, they report that noise enhances the sensitivity of these neurons, enabling them to respond optimally even to weak signals. Similarly, Basalyga and Salinas (2006) document the constructive role of random variation in neural-network performance in a simple classification task as well as in complex tasks requiring coordinate transformations. Moreover, Stein et al (2005) claim that random variation is even more important for the auditory system, which deals with high-frequency stimuli. They argue that random variation improves transmission fidelity for high-frequency signals.

Interestingly, McIntosh et al (2008) conducted a correlational study comparing the variability of the brain signal and behavioural variability. These were obtained respectively from neural (EEG) and behavioural data (response latency and accuracy scores) from 55 children (aged between 8 and 15) and 24 adults (aged between 20 and 33) performing face-recognition tasks. Results show that brain variability is negatively correlated with response-latency variability, while it is positively correlated with accuracy scores. That is, the greater the brain variability, the less the response variability and the higher the accuracy scores are. Brain variability is also found to increase with age. The researchers conclude that brain 'noise' seems to enlarge brain's "capacity for information processing" (p. 2). This conclusion agrees with speculations on the role of random variability in enhancing brain adaptability to uncertainties in the world (see e.g., Boly et al 2007). More generally, their results illustrate how random variability increases the efficiency of nonlinear systems including neural systems.

The implication for phonetic research is obvious. If random variability is essential for the functioning of complex systems and the human brain, as the above-mentioned studies suggest, then it might also be relevant to language activity. Let us not forget that within our field, researchers usually add random noise to synthetic speech to improve its acceptability by increasing its perceived naturalness (e.g., Portillo 2002; Wouters & Macon 2002). In all likelihood, an investigation of phonetic variability that does not exclude random variations will

add to our understanding of the subject. Such an investigation might also improve our chances of resolving one of the infamous paradoxes in phonetic research. Here is how Jeffrey Elman (1992) portrays the paradox:

It is a curious paradox that some of the tasks that humans carry out with the least conscious awareness and with the greatest facility are precisely those tasks that seem to be the most complex and have been most resistant to analysis. The acoustic/phonetic processing of speech is one such domain. Introspection yields little insight into how this processing is done, and most listeners fail even to recognize that the task might be difficult. Yet attempts to duplicate this processing in machines have not been very successful (p. 227).

Needless to say, addressing adequately the question of phonetic variability holds one of the keys to understanding phonetic processing and subsequently resolving the above paradox. For these reasons, I wish to extend the applicability of the term 'lawful variability' to cover variation that is currently described as random for the simple reason that it defies explanation. To me, lawfulness does not presuppose predictability or explicability. Any variation that is not due to some identifiable performance-related errors on the part of speakers or on the part of researchers, but for which we still have no explanation, is just as lawful as variation whose sources we have already identified.

Figure 6-2 below depicts the territories of lawful and non-random variations according to the Variability Field approach, which I sketch next.

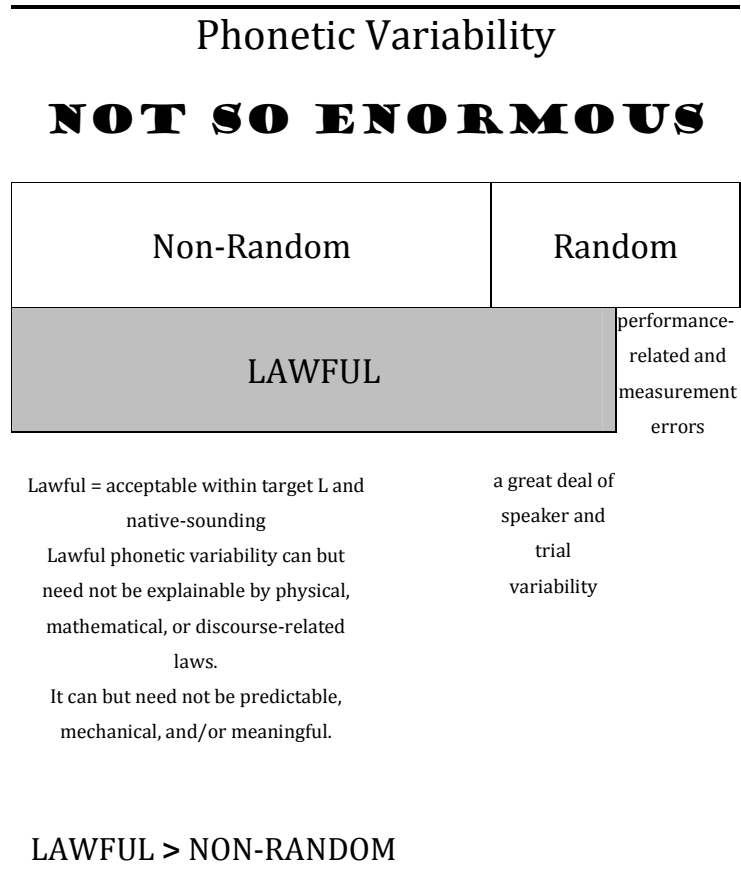


Figure 6-2: Phonetic Variability according to VFM: Lawful > Non-random

6.3 Phonetic Variability as Fields: The Variability Field Model

6.3.1 Overview

A central presumption of VFM is that phonetic variability is limited and structured in such a way that it can be 'fielded'. Hence the name variability fields. There are bounds to how widely the phonetic rendition of a unit of speech can vary and still be recognised as an acceptable and native-sounding rendition of that unit. There exists by now ample evidence supporting the conclusion that physiological, phonological, prosodic, morphological, lexical, and discourse-related factors have a constraining impact on phonetic variability (see e.g., Baese & Goldrick 2006; Baese et al 2007; Campos-Astorkiza 2007; Johnson 2001; Lavoie 2002; Lindblom 1986; Manuel 1990; Meunier et al 2006; Tabain and Perrier 2007, 2005; Vaux & Samuels

2005). To take but one example, the system of contrast in a given language limits how much variation contrasting sounds can exhibit. Figure 6-3 reproduces a schematic from Vaux and Samuels (2005: 411) graphing the constraining effect of a three-way laryngeal contrast on the amount of VOT variation in the realisation of the contrasting sounds. As is shown in the figure, plain voiceless series (T) exhibits the least variation while the other two series (D) and (T^h) have larger latitudes of variation in the lead and lag directions, respectively. The schematic actually rests on real VOT data from Nepali, which has a four-way laryngeal contrast (for more on this, see Poon & Mateer 1985; Vaux & Samuels 2005). See also Johnson (2001) for a similar argument using comparative data from English and Korean.

COPYRIGHT MATERIAL

Figure 6-3: A schematic graphing the constraining effect of a three-way laryngeal contrast on the amount of VOT variation in the realisation of the contrasting sounds (reproduced from Vaux & Samuels 2005: 411)

Underlying VFM is a hybrid, integrative approach to phonetic representation and processing, placing variability where it belongs—at the heart of phonetic data. The hybridity of the model comes from its combining abstractionist and episodic schemes. Specifically, VFM recognises as component fields both a quasi-episodic phone-field and an abstract phoneme-field.⁹³

However, these fields are not a complete innovation of VFM. They actually share certain features with prototype theory (e.g., Kuhl 1991; Kuhl & Iverson 1995; Samuel 1982) and exemplar-based approaches to the lexicon (e.g., Goldinger 1998,

⁹³ I do not attempt to define these representational units rigorously. Rather, I use the terms for descriptive convenience to code a representational distinction. Roughly, as a field label, the phone describes a context-specific realisation of a given sound (cf. Ladd 2006). Beyond that, it refers to each perceptibly different instance of that sound (cf. Laver 1994) which can join the potential population of a phone-field. In contrast, a phoneme here is a grouping label attached to a collection of phone-fields that share certain characteristics.

1996; Johnson & Mullennix 1997; Palmeri et al 1993; Pierrehumbert 2002, 2001). For example, the phone-field can be likened to a very tiny sub-space of an exemplar cloud emptied of all inhabitant exemplars except those occupying the modal interval. As such, a phone-field has an internal structure within a bounded spread area (see §6.3.3.1 for more details). To some extent, this construction is reminiscent of prototypical configurations of phonetic categories which distinguish between central and boundary tokens (e.g., Allen & Miller 2001; Kuhl 1991; Kuhl & Iverson 1995; Samuel 1982). Moreover, the definition of prototype found in neural research which appeals to a sound's frequency of occurrence (e.g., Guenther & Gjaja 1996; Näätänen et al 1997) is compatible with VFM's notion of the modal interval. See chapter five for details.

By the same token, the phoneme-field, being a constellation of phone-fields, is like a cloud of dispersed exemplars all carrying a single phoneme label⁹⁴ (see Pierrehumbert 2001). Needless to say, exemplars populating a given phoneme cloud considerably outnumber the phone-fields within a phoneme-field. Accordingly, the phoneme-field is a lot more coarse-grained than a phoneme cloud of exemplars. See §6.3.3.2 for more details.

On the other hand, the phoneme-field borrows from prototype theory the common definition of prototypicality, which emphasises well-articulatedness and peripherality over typicality (Lotto et al 2000). A prototype, thus defined, is actually not bound to a context and, as Lotto et al (1998) observe, might remain a meta-linguistic judgement. In other words, a prototype in this sense might not exist in real speech situations. As we will see later, VFM allows the construction of a meta-phone—an idealised, context-free pseudo-phone-field within a phoneme-field which serves certain metalinguistic purposes. Thus, the meta-phone is prototypical in the sense described above. See §6.3.3.2 for details.

⁹⁴ There is no agreement among those who actively subscribe to the exemplar-based approach regarding what unit of speech a cloud of exemplars represents. There are at least three candidates in the literature: the word (e.g., Johnson 2007), the phoneme (e.g., Pierrehumbert 2001), and non-automatic allomorphs (Välimaa-Blum 2009). See also Hall (2008) for one possible exemplar-based conceptualisation that explicitly integrates allophonic contributions.

VFM is also integrative in that it offers an inclusive treatment of linguistic, paralinguistic and extralinguistic effects.⁹⁵ Undoubtedly, there is a multitude of variables contributing uniquely or collectively to these effects that one might wish to model. However, as I am more concerned here with laying out the overall architecture and conceptual framework of the model, I will limit the discussion and illustration of these effects to a tiny subset of variables. Specifically, the linguistic effects I will attempt to model are phones and phonemes in the sense adopted in this thesis; the paralinguistic effects will include speaking rate and orthography; and the extralinguistic effect will be speaker gender. Figure 6-4 summarises these effects in a flowchart. The principles emerging from this limited-scope discussion should hopefully apply, probably with some modification, to other effects not modelled here.

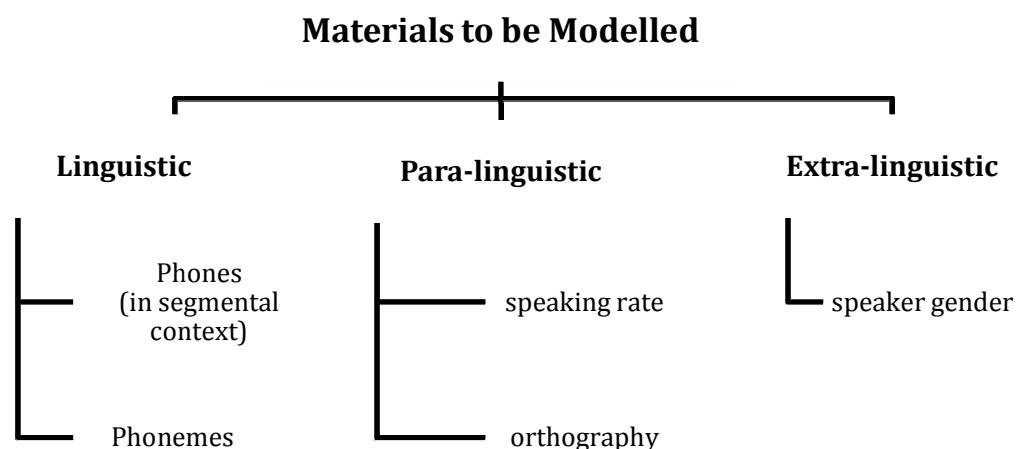


Figure 6-4: Overview of linguistic, para-linguistic, and extra-linguistic effects modelled in terms of VFM

It is important to note that VFM does not only take a classificatory stand on these effects, merely specifying where they belong in the model. VFM is also concerned with determining what and how they contribute to our understanding and modelling of phonetic phenomena. Put differently, VFM defines roles underpinning the nature of the contributions of these effects. Among these are (1) the capacity to form fields, (2) the capacity to provide discrete or continuous background against

⁹⁵ See Laver (1994) and Traunmüller (1994) for definitions and illustrations of these effects. See also Lachs et al (2003) for a different view.

which a field is accessed, and (3) the capacity to exert an attraction force pulling the realisation and/or accessing of the material within a phone-field in different directions. These roles are summarised in Figure 6-5. The attraction forces can differ in temporality and location, with some being momentary and others being lingering. Also, the domain of effect of these forces can be localised within certain phoneme-fields and/or limited to specific phones. These effects can also be global, extending to all phones within a given utterance.

VFM treats these effects differentially. For example, only linguistic material may form fields, whereas paralinguistic and extralinguistic effects may provide a background and/or exert an attraction influence. More specifically, phones and phonemes will form fields.⁹⁶ These fields will be accessed within a background defined in terms of speaker-gender or in terms of speaking rate. Finally, orthography will be depicted as exerting an attraction force. I present a detailed sketch of the model in §6.3.3 and §6.3.4.

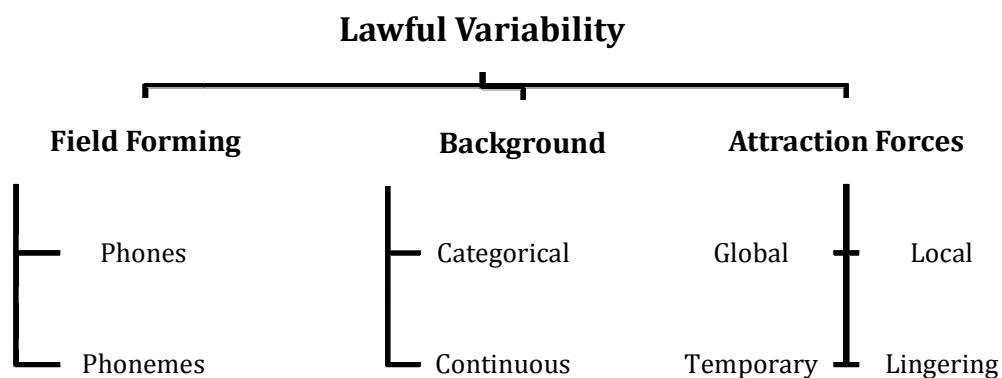


Figure 6-5: Lawful variability in VFM as forming fields or backgrounds, or as exerting attraction forces on the realisation of the material within a phone-field

6.3.2 Structure in Variability

The Variability Field approach starts with this simple observation. A lot of phonetic variations are below the threshold of perceptibility. Thus, they need not show up in our measurements nor contribute to our representation and assessment of phonetic variability. This would be our first step toward finding structure in

⁹⁶ In principle, larger units of speech like morphemes and words can form their own fields. I do not investigate this possibility here.

variability. To illustrate, I have plotted in Figure 6-6 a hypothetical⁹⁷ set of normally distributed duration data. The symmetrical shape and values of the data are for illustrative purposes. As we can see, there are 50 physically different data points. That is, duration here varies along a range covering arbitrarily-distanced 50 different values. To explore the same dataset within VFM, we need to examine the data in terms of intervals rather than as individual data points. To visualise data in intervals, we first need to fix interval width. For the purposes of illustration, I have set bin width to 20ms, which, as we will see in the next chapter, is a plausible bin width reflecting a jnd of 20% of a reference vowel whose mean duration is 100ms. This gives us only five perceptually different data intervals, indicating that duration varies along a range of five possible values. Importantly, the VFM-based exploration enables us to see for ourselves that phonetic variability is even constrained by our own perception. At the same time, and perhaps more importantly, this way of exploration reveals a kind of structure in the dataset. Looking at the data display in Figure 6-6, we can locate, at a quick glance, the modal interval, tell the number of intervals, and form an idea about the proportion of the data within the modal interval relative to the number of the intervals in the histogram.

⁹⁷ The use of hypothetical data is for simplified exemplification. In the next chapter, I will examine real data illustrating the points argued for here.

50 53 60 64 69 71 74 75 78 79 83 84 86 87 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 111 114 115 118 121 122 124 125 127 129 133 135 140 146 150

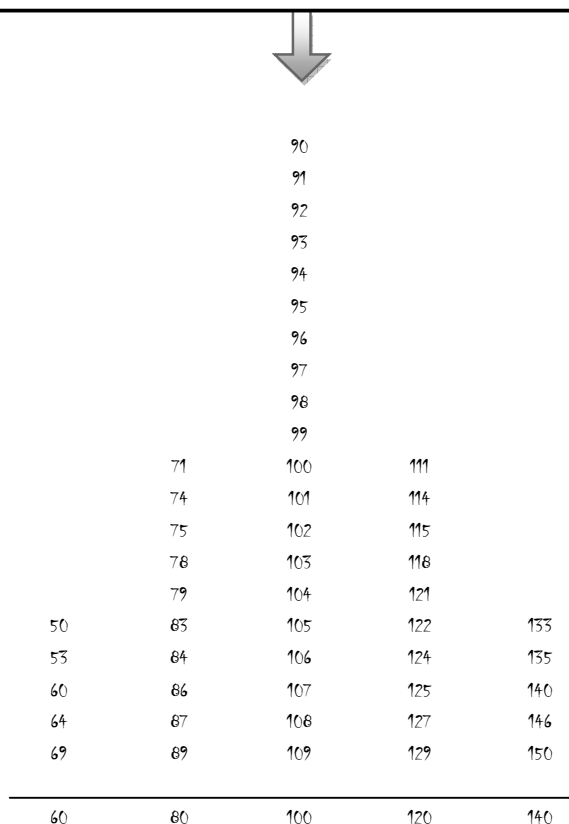


Figure 6-6: A hypothetical set of normally distributed duration data ($n=50$) arranged horizontally from smallest to largest (top) and plotted as a histogram with a bin size of 20ms (bottom)

Another step we can take towards finding structure in variability is to recognise that a great deal of the phonetic variability that we subject to analysis is actually composite variability (i.e., attributable to a variety of causes). This variability can be further analysed so that variations due to disparate sources are separated.

Theoretically, the analysis can continue up to a point where a certain portion of variability is no longer analysable given our current knowledge and resources, which are, admittedly, not unlimited. I will refer to this non-analysable portion of variation as primary variability. Now it must be noted that both primary and composite variations are lawful as long as they remain native-like. That is, only variations that native users of the language in question find inappropriate, be it relative to a given context or in absolute terms, for known or unknown reasons, should be outlawed. Granted, composite variability will be less random by virtue of the traceability of its causes, which are responsible for its composite character.

What remains after the extraction of the relevant sources of variation is primary variability, which as such can be seen as random. Note that, strictly speaking, many instances of what our limited means compel us to treat as primary can still be instances of incompletely analysed composite variability. A great deal has been written on sources of variability (e.g., Benzeghiba et al 2007; Klatt 1986, 1976; Labov 1986) but very little, if any, on the primary-composite distinction suggested in this thesis. That said, this chapter contributes towards filling this gap in the literature, with a more specific goal—to reframe our discussion of phonetic variability.

Regarding composite variability, VFM is initially concerned with identifying the various sources of variation, with a view to extracting the variations they cause in order to arrive at primary variability. These sources, of course, include linguistic and non-linguistic factors. Once we arrive at primary variability, the focus will be on its distributional properties such as modality, spread, and shape. These can be defined in relation to the notion of the modal interval and the range. As I explained in the previous chapter, the modal interval is the highest bar in a frequency histogram; it is where data dense up. The number of modal intervals in a dataset determines whether the dataset is unimodal, bimodal, or multi-modal. The size of the range defines the spread of the dataset. Hence, we can have widely-dispersed versus narrowly-dispersed datasets. The location of the modal interval relative to the spread area defines the shape of the dataset. For example, depending on where the modal interval falls relative to the spread area, we can have central-gravitation, left-gravitation, and right-gravitation fields. Of course, as shown in the previous chapter, the gravitation area and spread of the field provide respectively the central tendency and dispersion statistics within VFM.

Figure 6-7 below summarises the relationships between composite and primary variability. The smoothed lines in the figure are only for simpler visualisation.⁹⁸ Later, I will use histograms exclusively. As is clear from the figure, adding a source

⁹⁸ Of course, given the central limit theorem, unless there is very little or no overlap among the relevant subsets of the data, a large set of sampled data like what we have in the figure will tend to approximate the normal, bell-shaped distribution when plotted as a frequency histogram. This happens irrespective of the underlying distribution in the population from which the data have been sampled. The distribution of the underlying population may still not be normal, though (Mark Huckvale, p.c. 2010). Therefore, it is very important that gravitation and spread are calculated for primary rather than composite variability. It must be noted in this respect that, unfortunately, the vast majority of published phonetic data are examples of composite variability.

of variation to a set representing primary variability can create composite variability, with the possibility that its basic distributional properties such as modality, spread, and shape change as a result. Conversely, primary variability can be derived from composite variability by simply removing the contribution of the relevant source.

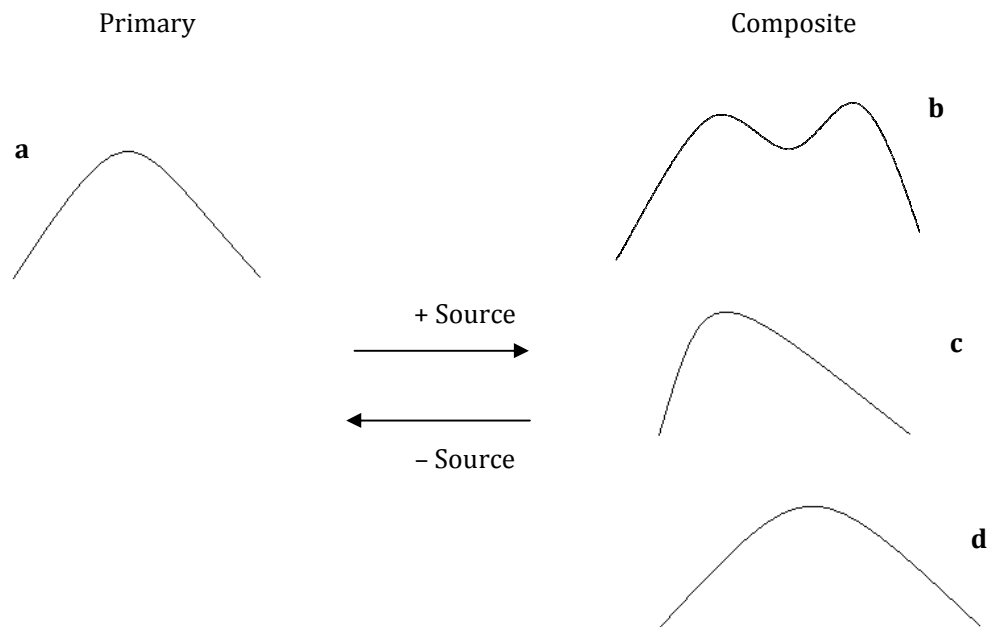


Figure 6-7: Schematics illustrating the bi-directional relationship between composite and primary variability; as indicated by the arrows, adding a source of variation to a set representing primary variability, as in (a), can create composite variability, as in (b-d); removing the contribution of a source of variation from composite variability can result in primary variability; distributional properties of a set of variability such as modality (b), shape (c), and spread (d) can be affected by the addition or removing of a source.

This bi-directional derivability of primary and composite variability is a manifestation of the structure resident in phonetic variability. To illustrate, let us consider a hypothetical F2 dataset collected from a single speaker as follows: 50 CV stimulus items (10Cs x 5Vs) each repeated 27 times at a normal speaking rate. The test vowels are [i], [e], [a], [o], and [u]; the consonants in the test items are [b], [d], [g], [v], [z], [ʃ], [m], [n], [r], and [w]. In Figure 6-8, six possible modes of analysis are schematised. In (A), all the 1350 data points are considered for

analysis; in (B), only vowels following labial consonants are considered, totalling 540 data points. Of these only two vowels are considered in (C): [a] and [i] with a total of 216 data points; in (D), only [a] data are considered making up 108 data points; in (E), the labials preceding [a] are restricted to only [b] and [m] giving 54 data points in total; and finally in (F), [a] following only [m] data, making up 27 data points, are plotted. Examining the variability in all six data plots in terms of VFM, we conclude that only variability in (F) is primary, while that in (A) through (E) is composite.

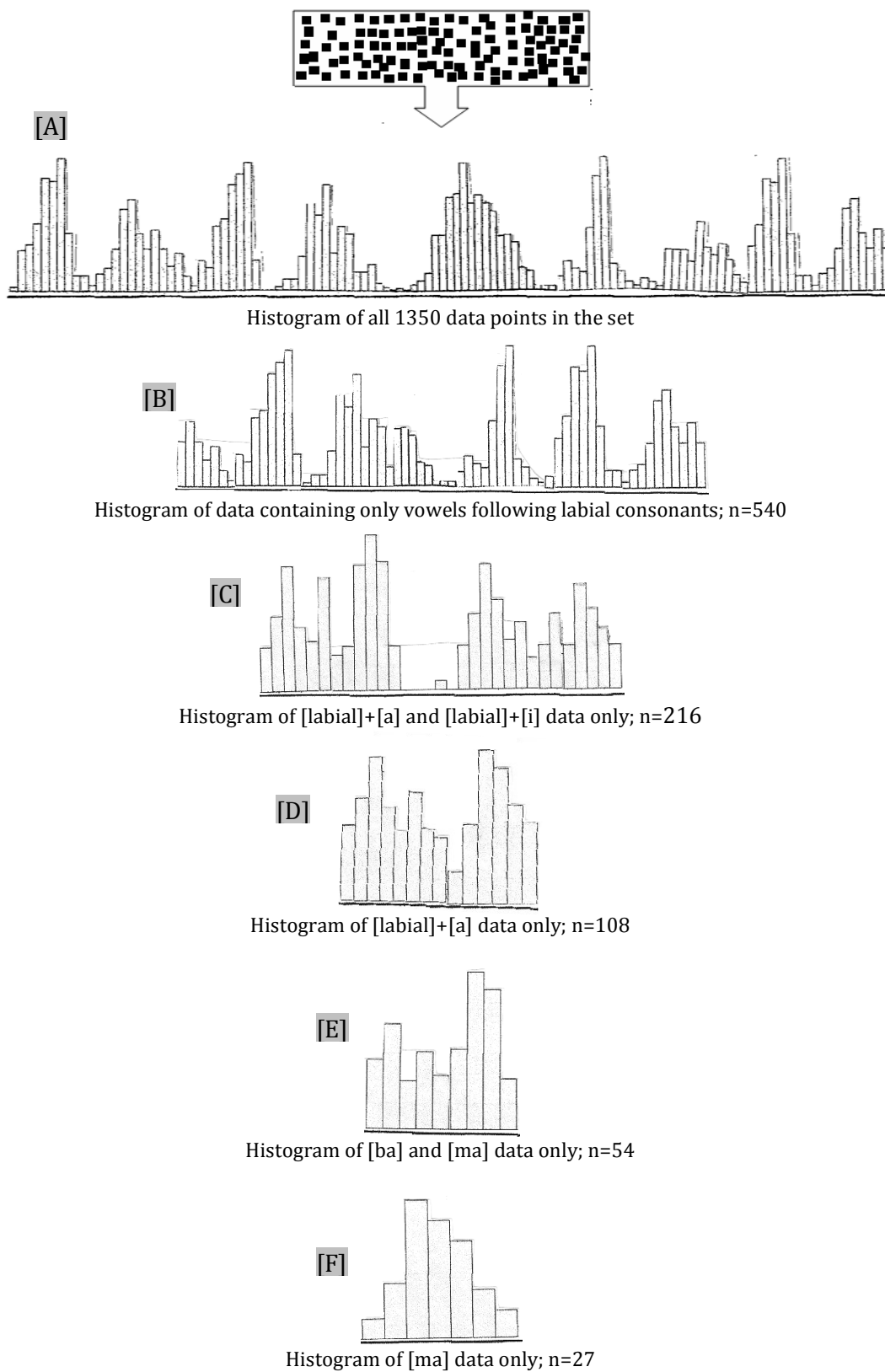
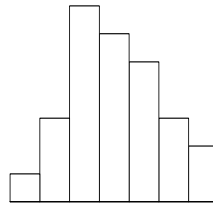


Figure 6-8: Histograms illustrating six possible modes of analysis of a hypothetical F2 dataset collected from a single speaker repeating 50 CV items 27 times; Vs= [i], [e], [a], [o], and [u]; Cs= [b], [d], [g], [v], [z], [ʃ], [m], [n], [r], and [w].

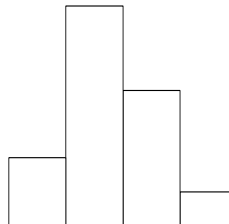
As I explained previously, for primary variability we need to locate the modal interval and the range. These make up the essential components of the phone-fields within the model. To do that, we need to make sure that the inter-midpoint distance represents a perceptible difference, e.g., by applying jnd-based binning. As it stands, bin width in histogram (F) is arbitrary. Therefore, I have re-plotted dataset (F) as (F-ii) below setting bin width to an appropriate jnd-based width in Hz (see chapter seven for more on this). As we can see, jnd-binning as in (F-ii) effects some pruning down of the data set in (F), which I have reproduced as (F-i) for the convenience of the reader. Finally, in (F-iii), I have isolated the modal interval and kept the spread line, which, as far as single-talker datasets are concerned, are all we need for a primary phone-field. See Figure 6-9.

[F-i]



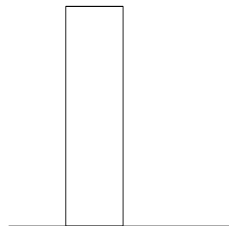
Histogram illustrating primary variability using an arbitrary bin width; perceptible and imperceptible differences are shown.

[F-ii]



Histogram illustrating primary variability using a jnd-based bin width; only perceptible differences are shown.

[F-iii]



Histogram illustrating the components of a phone-field: the modal interval and spread line

Figure 6-9: Histograms illustrating three possible analyses of [ma] data (n=27) plotted in Figure 6-8

However, in phonetic research, there are many researchers who are interested in explaining group behaviour rather than documenting the minutiae of an individual's phonetic behaviour. To these researchers, VFM offers a specific strategy for dealing with multi-talker datasets. However, before discussing that strategy, it is useful to familiarise ourselves with the basic architecture of the model, which I attempt next.

6.3.3 Field-Forming Variability

6.3.3.1 The Phone-Field

A phone in its immediate phonetic context forms a variability subfield, called here a phone-field. For reasons of space, I will only discuss and illustrate segmental context here. The exact number of sounds making up the context window will not have a direct bearing on the concept or functionality of the phone-field. For the purposes of this thesis, however, segmental context will not extend beyond two neighbouring sounds on both sides of the field's phone. This decision is certainly open to question. However, it is not entirely arbitrary but is actually inspired by experimental evidence suggesting that segmental effects diminish with distance. For example, coarticulatory effects are generally strongest in adjacent sounds (e.g., West 2000, 1999).⁹⁹ Examining the temporal extent of vowel-to-vowel coarticulation, Grosvald (2009) found that, as distance increased, the effect became inconsistent and only appeared in the production of some speakers of his sample, thus confirming the findings of Gay (1977). Importantly, the strongest coarticulation effect in the Grosvald study is between vowels across one consonant sound. Perception-wise, Grosvald reports that this coarticulatory effect was useable to all the listeners in his experiment. Beyond this distance, however, the coarticulation effect was barely perceptible.

On the other hand, numerous studies show that coarticulatory effects extend over a larger distance than that covered by the two-segment window in this thesis (see e.g., Hawkins & Slater 1994; West 2000, 1999). However, the formulation above serves the illustrative purposes of the thesis sufficiently for the present. Let me quote Kühnert and Nolan (1999: 28), who rightly point out that “truly quantitative

⁹⁹ In terms of directionality and domain, coarticulatory effects are not necessarily symmetrical (see Gay 1974 for articulatory data and Recasens 1987, 1989 for acoustic data). See also Bladon and Nolan (1977) and Magen (1997) for a different result.

statements about coarticulatory effects are yet difficult to derive". This is an area where we still need more research employing case-study as well as cross-language methodologies.

Undoubtedly, a superior definition of neighbouring environment is one that allows for more specificity with respect to segments, languages, and speakers. For example, a segment-sensitive criterion for defining the length of the coarticulatory window may be better than a criterion that applies indiscriminately as I have done in this thesis. There are reports in the literature supporting the conclusion that different sounds have different coarticulatory properties (see e.g., Bladon & Al-Bamerni 1976; Bladon & Nolan 1977; Recasens 1987, 1985). There are also reports in the literature that coarticulatory effects are language-specific (see e.g., Gobl & Ní Chasaide 1999; Boyce 1990; Manuel & Krakow 1984; for a review see Manuel 1999). Also important is the suggestion that speaker idiosyncrasies have a role to play in coarticulation effects (e.g., Nolan 1985, 1983). For the illustrative purposes of this thesis, however, I shall stick to the arbitrary criterion of limiting the segmental environment to only two segments on both sides of a field's phone. As I explained above, the exact length of the coarticulation window will not affect the concept or functionality of the phone-field as construed in this thesis.

To illustrate what a phone-field is, suppose we have a lexicon comprising only these words: 'ket', 'ketem' and 'meket'. There will be eleven different phone-fields amongst these three words, with 'ket' and 'ketem' sharing the [k]-field, and with 'ket' and 'meket' sharing the [t]-field. These are given in Figure 6-10. In each field, the sound in boldface is the field's phone.

[ket] has three different phone-fields:



[ketem] has five different phone-fields:



[meket] has five different phone-fields:



Figure 6-10: Phone-fields in a hypothetical three-word lexicon

As pointed out above, every phone-field has a measure of location and a measure of dispersion. These are, respectively, the modal interval, which stands for the field's gravitation, and the range of intervals, which represents the field's spread. By identifying the modal interval, we can determine (1) its frequency and (2) its numerical value, which can be the mid-point or the boundary values of the interval. So, (1) and (2) define, respectively, the force and the location of the gravitational area of the phone-field. Likewise, determining the spread of the field consists in identifying (3) the number of intervals and (4) the numerical value of the overall range of the dataset.

Accordingly, there are two count statistics here. These are a frequency count representing gravitation, and an interval count representing the size of spread. Together these counts define the variability magnitude of the field as expressed in terms of the Variability Field Index (VFI) introduced in the previous chapter. Roughly speaking, VFI is the ratio of spread to gravitation. As such, VFI is VFM's equivalent of the commonly used Coefficient of Variance (C_v). But unlike C_v , VFI only uses count data, and is thus independent of the numerical values of its components. See chapter five for details. Figure 6-11 is a flow chart showing the components of a phone-field and the relations holding among them.

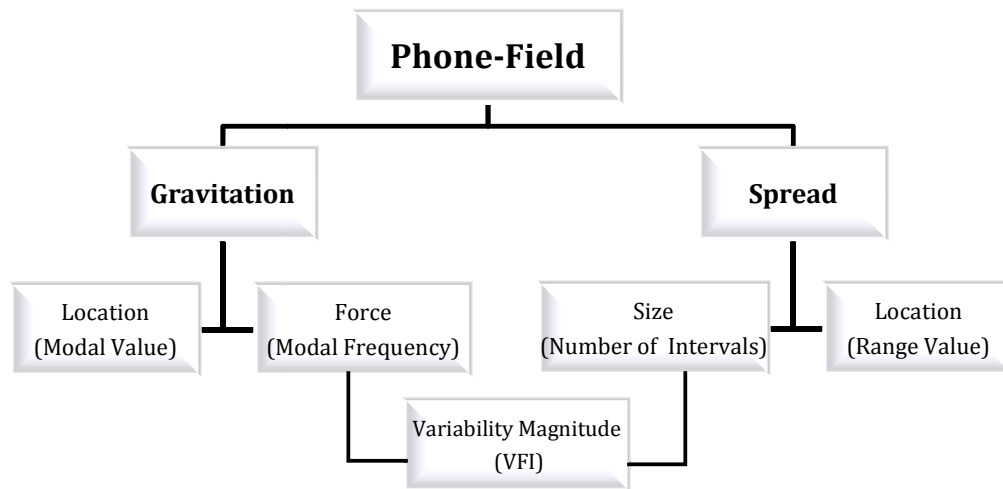


Figure 6-11: A flow chart showing the components of a phone-field

By the same token, we have two non-count summary statistics. These are the modal value and the range value. These define, respectively, the location of gravitation and the outer boundaries of the field. Taken together, these summary statistics make up a simplex co-ordinate system. This can be visualised as a line along which we calculate the distance between two given points, one of which is a point of reference. For our purposes, there are three candidate points of reference. These are marked on the line in Figure 6-12. These reference marks stand for the mode, the maximum, and the minimum. By way of illustration, I will consider the co-ordinates of 6 data points (A through F) shown in the figure. The co-ordinates of each will be the signed¹⁰⁰ distance separating that point from its closest reference point. As is clear from the figure, the reference point for both A and B is the mode, with A lying at a plus distance from the mode, whereas B is located at a minus distance from the mode. Similarly, C is at a minus distance from the maximum, while D is at a plus distance from the same reference point. Finally, E is at a plus

¹⁰⁰ Strictly speaking, attaching signs (+/-) to distance is not mathematically appropriate. Although distance can go in a positive or negative direction from a reference point, it is the direction that is positive or negative, not the distance itself (Panovska-Griffiths, p.c. 2010). The description above is only for simpler exposition.

distance from the minimum while F is at a minus distance from that reference point.

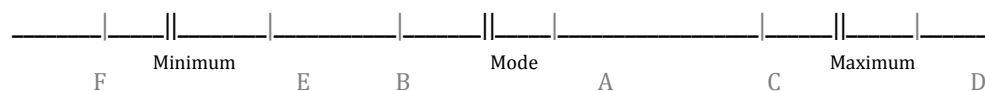


Figure 6-12: A simple co-ordinate system with three reference points and six different points lying at various distances from the reference points

Within VFM, the frequency of the modal interval can exert an inward attraction force, causing a shrinking effect on the spread area. This shrinking effect can be temporary or lingering, depending on the amount and type of the influence. An example of a temporary shrinking effect is what is known in the perception literature as feature saturation¹⁰¹ (see Eimas & Corbit 1973; Diehl & Kluender 1987). The effect is usually brought about in a three-phase perception experiment. Roughly, the experiment proceeds as follows: during the first phase, a category boundary in a stimulus continuum along a phonetic parameter, e.g., VOT for a /b/-/p/ contrast, is located using responses from a group of participants. Next comes the saturation phase, where participants are exposed to a stream of repetitions of an endpoint stimulus, e.g., [ba]. In the third phase, participants re-take the same perception test as in phase one, and the category boundary is re-located. An important finding of this paradigm is that the boundary shifts between the test and the re-test phases towards the category to which the saturating stimulus belongs.

According to VFM, boundary shifting in response to a saturating stimulus as brought about in the laboratory is a manifestation of a temporary shift in the bounds of the phone-field. It is an inward shrinking effect caused by an increase in the gravitational force of the field. In real-life situations, on the other hand, the phenomenon is also possible and may be instantiated as sound change or adaptation. At the same time, VFM allows the possibility for the outer limits of the field to get stretched to take on new members. Again this can be short- or long-

¹⁰¹ Feature saturation is more commonly referred to as selective adaptation (e.g., Repp & Liberman 1987; Miller et al 1983). I use the less common term 'feature saturation' because it fits the theme of the chapter better.

lived. Indeed, part of the dynamicity of a phone-field lies in the updateability of its gravitation and spread components. However, to appreciate how this is possible, it is instructive to learn first how a phone-field is constructed.

The construction of a phone-field involves the formation of a gravitational area within a spread area, which defines the field's outer bounds. The requirements for setting up the gravitational area are different from those needed for establishing the spread area. Specifically, gravitation, being a property of the modal interval, requires, by definition, frequent repetitions of perceptually non-distinguishable tokens of the relevant phone in its relevant context. In contrast, spread, which represents dispersion, requires an element of non-uniformity, or even diversity. By definition, then, there is a need for perceptually distinguishable tokens of the relevant phone in its relevant context. Of course one way to ensure diversity is to vary production conditions. Since we are dealing here with phone-fields, which should ideally exhibit primary variability, diversification must not affect segmental context, but be, instead, confined to paralinguistic and extralinguistic effects.

As I explained before, the path from composite variability to primary variability mostly consists in reducing the analysability of the relevant dataset. Within the current version of VFM, only variability related to phonetic context (i.e., allophonic variation) is decomposed before the formation of a phone-field. Within a given phone-field, then, there is always only one segmental context. Variations due to background sources such as speaking rate or speaker attributes like gender (i.e., indexical variation) will be decomposed before the accessing of the phone-field within that background. So, in principle, these can contribute to the initial set-up of a phone-field, especially that they will be factored out later on. I give more details in §6.3.4, where I discuss background effects.

By extension, it seems reasonable to conclude that the construction of a phone-field involves attending to repetitions of the phone in its phonetic context in different backgrounds. On this view, then, repetitions of a newly encountered [aba], for example, produced by the same speaker under more or less the same conditions will be enough to form a gravitation area and modal frequency, but these repetitions will not be enough for establishing a spread area.

Similarly, being exposed for the first time to stimuli such as [aba], [apa], and [ada] all produced by the same speaker under more or less the same conditions and with no repetitions will not lead to the creation of gravitation nor spread for three different sounds. In other words, no phone-field will be constructed on the basis of such data. It is only when a learner hears repetitions of [aba], for example, by different speakers in different settings that a phone-field is constructed with both gravitation and spread.

More generally, learning a language, according to VFM, involves, in part, the construction and maintenance of the appropriate phone-fields.¹⁰² Maintaining a phone-field includes assessing the relevance of the incoming stimuli and updating accordingly the gravitation and spread areas of the field, where appropriate. This layout not only allows for the developmental errors which are in abundance during the early stages of L1 and L2 acquisition. It also benefits from these developmental errors. The phone-field in VFM is not a static, idle construction but is actually dynamic, developmental, and data-driven.

6.3.3.2 The Phoneme-Field

The construction of a phoneme-field involves grouping the relevant phone-fields into a super-field. For the hypothetical three-word lexicon presented in §6.3.3.1, there are four phoneme-fields, schematised in Figure 6-13. In addition, each of these phoneme-fields can contain an idealised pseudo-phone-field, called here a meta-phone¹⁰³. The meta-phone is context-free and can act as a source of attraction affecting different phone-fields within a phoneme-field. The construction of the meta-phone can be sensitive to orthography (cf. Taft 2006; Taft & Hambly 1985). The meta-phone also plays an important role in dealing with unfamiliar sequences including non-words (cf. Morais & Kolinsky 1995). There are numerous studies suggesting that listeners treat non-words and sounds within non-words differently from real words and sounds within real words (e.g., Ganong 1980; Whalen et al 1997). For example, Utman et al (2000) report a reduction in priming caused by manipulations in the acoustics of real-word primes. By contrast, with equivalently manipulated non-word primes, the researchers found no

¹⁰² I do not attempt to delve into learning here but observe that VFM-based inquiries into language acquisition introduce a new perspective on the subject.

¹⁰³ The nature of this idealisation is not central to the claims I make here. Representationally, a meta-phone can be made up of features (e.g., Halle 2002; Stevens 2002), elements (e.g., Harris & Lindsey 1995, 2000), or gestures (e.g., Browman & Goldstein 1989).

reduction. Similarly, Marslen-Wilson and Warren (1994) found that mismatching-coarticulatory cues had a disruptive effect on the processing of real words but not on the processing of non-words. A plausible conclusion of these and similar studies is that the processing of real words involves accessing context-bound phone-fields, which are determined on the basis of allophonic variations. In contrast, the processing of non-words involves accessing meta-phones, which are a-contextual idealisations where allophonic variations have no status.

Being context-free, a meta-phoneme can be called upon for processing, independently of indexical variations. The meta-phoneme here is VFM's version of what Luce and McLennan (2005: 601) call the "abstract codes [that are] untainted by surface variation". These codes, the authors claim, are utilised in speeded recognition tasks where the advantage of processing speech repeated in the same voice as in the training task disappears.

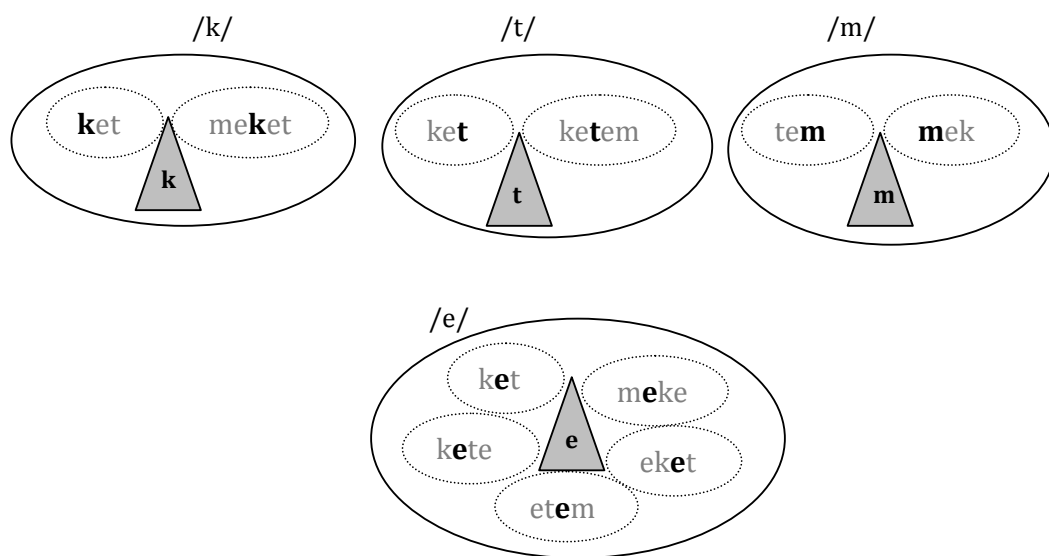


Figure 6-13: Phoneme-fields of a hypothetical three-word lexicon; phone-fields appear as small ovals inside a phoneme-field; meta-phones appear as triangles.

The construction of a phoneme-field is essentially a labelling activity that requires some form of abstraction. Elucidating the nature and mechanism of such an activity is beyond the scope of this thesis and will have to await future research. However, it should be remembered that labelling and abstraction may be both

involved in one of the cognitive skills that people innately have, and which many researchers marvel at—categorisation (see e.g., Duch 1996; Sedlmeier et al 2002). Interestingly, Lacerda (1997) observes that labelling occurs for supervised as well as unsupervised forms of learning. Moreover, there is by now a sizable body of empirical data providing evidence for the representation of the phoneme (see e.g., Cho & McQueen 2006; Gaskell et al 2008; Kazanina et al 2006; Ohala 1983; Ohala & Feder 1994). For example, McQueen et al (2006) provide behavioural data supporting the view that lexical processing requires phonological abstraction before lexical access. Furthermore, Näätänen (2001) reviews a large body of neural data pointing in the direction of the existence of an abstract phoneme trace in the human brain (see also Doufour et al 2010, 2007; Eultiz & Lahiri 2004; Lahiri & Marslen-Wilson 1991; Lahiri & Reetz 2002).

The skeletal description of the phoneme-field above is useful for visualising how language users come to possess and access an impressive knowledge of statistical and distributional properties of the sounds and sound patterns of their language. For example, the schematic in Figure 6-13 offers a simple way to estimate the comparative as well as absolute rate of prevalence of a given phoneme in the lexicon. Specifically, the larger the number of phone-fields within a given phoneme-field, the larger the number of the hosting environments, and the fewer the co-occurrence restrictions are.

More interestingly, the quasi-episodic nature of the phone-field, with frequency already inherent in the definition of the modal interval, makes it possible for speakers/hearers to estimate token frequency. However, given that not all tokens have traces in the phone-field, this subjective token-frequency estimate is expected to be attenuated downwards. In chapter five, I surveyed studies claiming that people only use a subset of the available data points (Kareev et al 2002; Peterson & Beach 1967). My call for emphasising data falling within the most frequent interval and de-emphasising data falling in the other intervals is in complete harmony with this observation. Accordingly, VFM makes the following prediction: subjective estimation of type-frequency should reflect closely objective estimations. The rationale behind this is that, unlike in the case of token frequency, language users have a detailed record of the segmental sequences that are permitted and actuated

in their language, which are called upon for estimating type-frequency. This prediction awaits examination.

These statistical and distributional effects are commonly accounted for by appealing to some form of probabilistic retrieval mechanism that the speaker/hearer is assumed to somehow possess and implement. By way of contrast, VFM attributes frequency estimation to a more primitive form of statistical thinking—the processing of untransformed count data. As I pointed out in chapter five, frequency-based Bayesian reasoning is cognitively superior to probability reasoning. See chapter five for details and references.

What the schematic above does not show yet is that two or more phoneme-fields can share one phone-field. Put differently, a phone-field may belong to different phoneme-fields, and as a consequence, a phoneme-field may contain disparate phone-fields. This is, in a nutshell, the essence of how VFM conceptualises and models contrast neutralisation. To illustrate, let us take a /t/-/d/ contrast that is neutralised word-finally. As shown in Figure 6-14, VFM captures both the contrast and its neutralisation by allowing the phoneme-fields for /t/ and /d/ to share the same phone-field, here **xx**t. Note that it is also possible that the phonetic context (**xx**) in the phone-field occurs in only one phoneme-field, say, the /d/-field.¹⁰⁴ This latter case is particularly true when no minimal or near-minimal pairs contrasting final /d/ and /t/ exist. Nonetheless, in both cases, neutralisation can be phonetically incomplete when the gravitation area within the relevant phone-field is pulled towards the meta-phone of the phoneme-field, resulting in a difference between the **[t]**-field gravitation when the /t/-field is accessed and its gravitation when the /d/-field is accessed. This difference can persist and give rise to bimodality. Meta-phone attraction can be intensified by orthography. In this sense, incomplete neutralisation is more likely for contrasts that are orthographically represented. Conversely, complete neutralisation is more likely for contrasts that are not orthographically represented.

¹⁰⁴ Recall that if word-final [t] consistently belongs to only one phoneme-field, we no longer have a case of neutralisation, but rather a static lack of contrast at the lexical level.

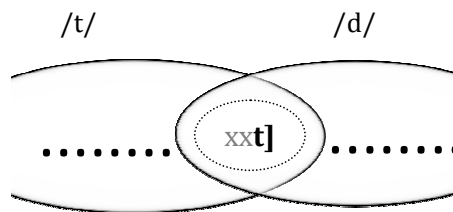


Figure 6-14: A schematic of a word-finally neutralised /t/-/d/ contrast

This relationship between phone-fields and phoneme-fields is a simplified manifestation of the multi-faceted and intricate relationship between phonetics and phonology. To illustrate another dimension of this relationship, consider the following schematics in the typology of vowel reduction. In language A, /a/ reduces to [ə] in X-context. For the sake of simplicity, I abstract away from the exact nature of the context. In language B, /a/ does not reduce to [ə] in that context but is realised as [a]. In language C, /a/ sometimes reduces to [ə] and sometimes it does not, independently of speaking rate, style, and register. Both pronunciations are acceptable in exactly the same context. Finally, in language D, /a/ reduces to [ə] only in fast speech. (50) lists how VFM deals with these facts. They are also graphed in Figure 6-15 below.

- (50) Typological facts of vowel reduction according to VFM
- a. Language A: xəx phone-field in /a/-field
 - b. Language B: xax phone-field in /a/-field
 - c. Language C: xəx phone-field and xax phone-field in /a/-field
 - d. Language D: bimodal xəx/xax phone-field in /a/-field

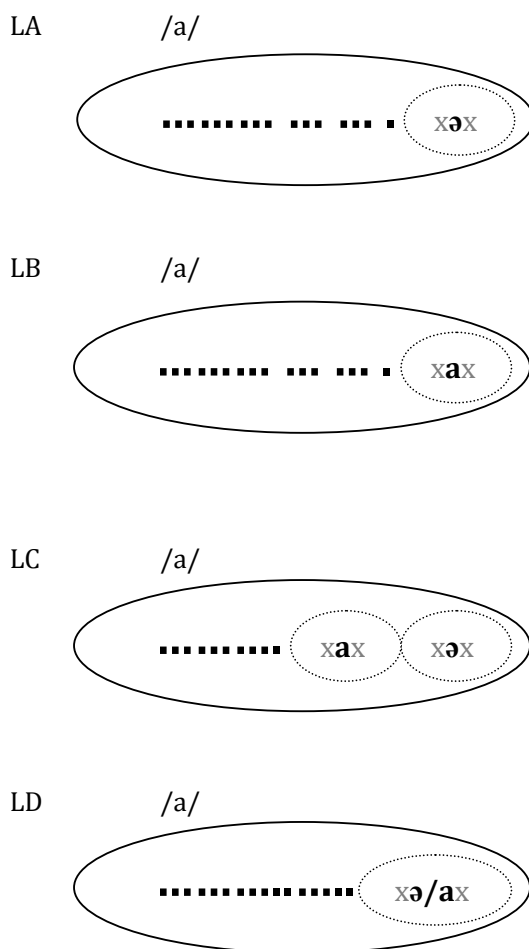


Figure 6-15: A schematic of vowel-reduction patterns in four hypothetical languages (LA-LD) according to VFM

6.3.4 Background-Forming Variability

VFM's position on variability that is due to paralinguistic and extralinguistic factors such as speaking rate and speaker-related effects like speaker gender is that this type of variability provides background to the various fields in the model. As backgrounds, speaking rate and speaker gender, which I attempt to model here, differ in important respects. Specifically, speaker gender is a binary variable whose categories are naturally discrete and mutually exclusive. Having said that, I must hasten to add that there is always the possibility of overlap in the physical values of the phonetic data produced by males and females. Figure 6-16 depicts four scenarios for positioning this binary variable in the continuous space of phonetic data. Specifically, in (a) there is a complete lack of overlap between the data produced by males and females; in (b), there is complete overlap which is visually

shown in the perfect alignment of the two boxes representing male and female data; and in (c) and (d), there is partial overlap in the physical values of the male and female data.

In contrast, speaking rate is a continuous variable which can, nonetheless, be artificially categorised into fast, normal, and slow rates, as graphed in Figure 6-17. However, coding data in this way will always have an element of relativity and arbitrariness. Deciding on what counts as fast or normal or even slow speech usually involves a comparison to a reference. It can rarely be decided in absolute terms. Likewise, deciding on exactly where to place the divide between, for example, fast and normal speech remains arbitrary in most cases.

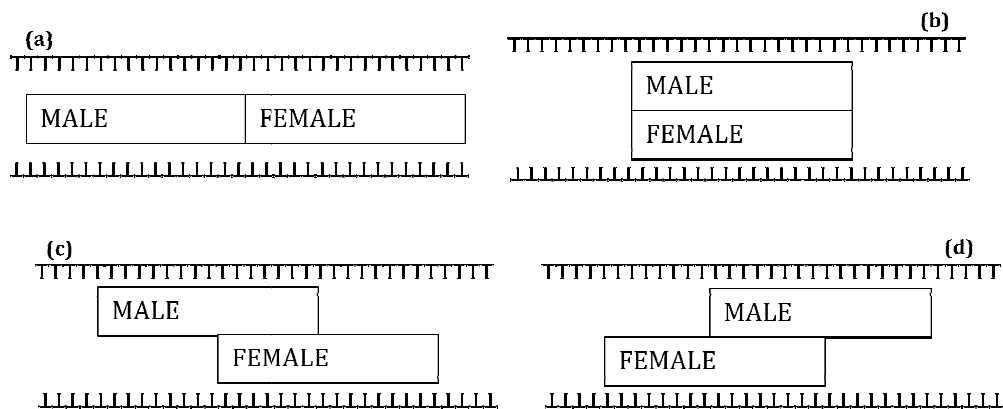


Figure 6-16: Four schematics of the binary variable 'gender' as background

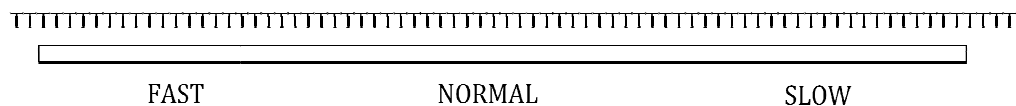


Figure 6-17: A schematic of the continuous variable 'speaking rate' as background

More concretely, according to VFM, if the incoming data along a particular acoustic dimension form a uni-modal frequency distribution (e.g., vowel duration by men and women), the phone-field will be set up with a single gravitational area. If the data form a bimodal distribution (e.g., vowel duration in conditions of fast and slow speaking rates), the phone-field will be set up such that there are two rate-

dependent gravitational areas. At any point in time, only one area of these is activated. Note that this configuration may solve the modality issue, but it will exaggerate the field's spread, which will then be the combined areas of both rate conditions taken together. It will also obscure any existing differences in the dispersion of these rate conditions. One way to deal with that problem is through autonomous layered-variability fields. Here each layer has its own modal interval and spread already defined. Each layer is accessed in its appropriate background. At any point in time, only one layer is activated. See graphs (a) and (b) in Figure 6-18.

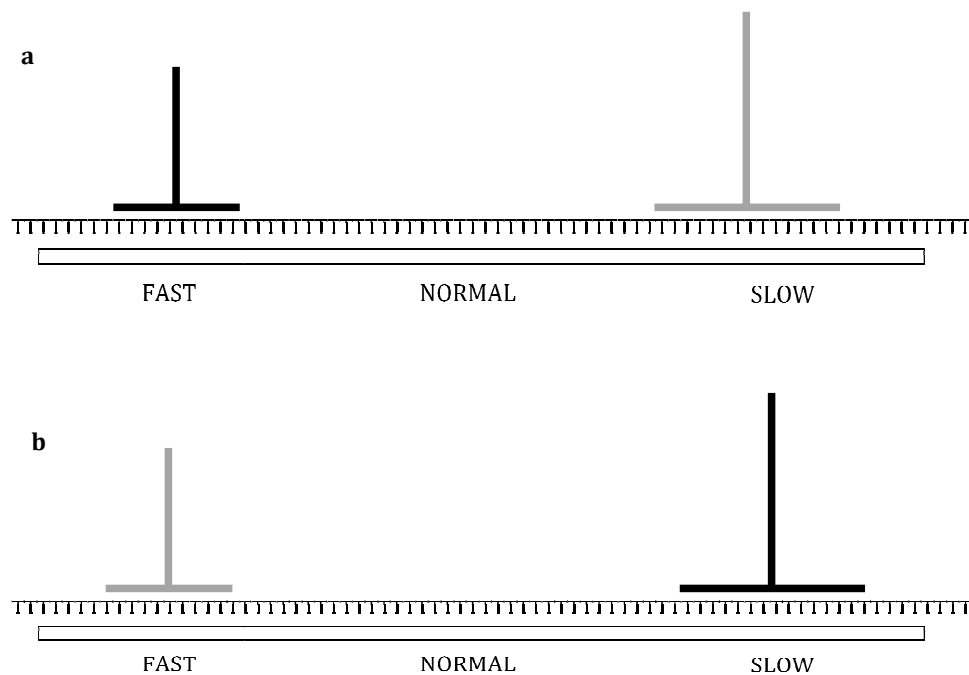


Figure 6-18: Autonomous layered-variability fields drawn against speaking rate as background; each layer has its own modal interval and spread already defined; at any point in time, only one layer is activated. The activated layer appears in black.

According to VFM, backgrounds are accessed before phone-fields. Phone-fields are processed against the activated background. Regarding speaker variability, this configuration captures the empirical observation that we attend to the voice of the speaker before we attend to the auditory message. With unfamiliar voices, it is usually the case that, for the first few milliseconds, we do not get what is being said. According to Mullennix et al (1989), voice information is processed very early. In fact, Mullennix and Pisoni (1990: 388), based on empirical data, conclude

that “the analysis of phonetic information contained in word-initial consonants is more dependent on the prior or concurrent analysis of voice information than vice versa”. The researchers also observe that voice information and segmental information are processed qualitatively differently. For example, they report that

[s]ubjects apparently can attend to dimensions of voice and selectively ignore irrelevant variation in the words. However, they have much more difficulty attending to words when there is simultaneously irrelevant variation in the voice of the talker (ibid: 389).

VFM treats these sources of information differentially, granting one the capacity to form fields, and the other the capacity to provide background to these fields. This arrangement can actually resolve one of the long-standing paradoxes involving the seemingly contradictory role of speaker variability in word-learning and in word-recognition. It is well-known that speaker variability can have a facilitatory effect on word-learning, but it can have an inhibitory effect on word-recognition. A number of researchers including Barcroft and Sommers (2005), Houston and Jusczyk (2000), Lively et al (1993), and Rost and McMurray (2009) have come to the conclusion that speaker variability is essential in learning. Specifically, they show that participants who are exposed to a stimulus set produced by multiple speakers learn and retain the target contrast much better than participants who are exposed to the same training set spoken by a single talker. In contrast, word-recognition is better in single-speaker conditions than in multi-speaker conditions (e.g., Bradlow & Pisoni 1999; Goldinger 1996; Mullennix et al 1989; Nygaard et al 1994; Sommers et al 1994). More intriguingly, McMichael (1999) found that, in the test phase of a sentence recognition paradigm, subjects responded ‘old’ to entirely new sentences when they were produced by a voice they had heard in the study phase. Sentences were more correctly recognised as ‘old’ when they were presented in the same voice as in the study phase than in a new voice. Similarly, Cole et al (1974) report that participants took a longer time to give same-different responses to stimulus pairs whose members were spoken in different voices than those spoken in the same voice. See also Palmeri et al (1993) for similar findings.

According to VFM, learning involves setting up and updating the necessary fields. Speech recognition involves accessing an already formed field. The processing of single-talker data requires very little background shifting. Interestingly, Allen and

Miller (2004: 3182) observe that “exposure to only a few different words [... may be] sufficient for a listener to learn a talker’s implementation of a given phonetically relevant acoustic property” (cf. Johnson 1997). This indicates that very little is required in the case of single-talker data for fields’ co-ordinates to be calculated and/or predicted.

As to data from different speakers, there will always be a renewed need for more input for the calculation of the fields’ co-ordinates with every new speaker heard. This may involve a lot of background shifting, which will in turn slow processing and may even hinder speech recognition (cf. Goldinger 1997). Interestingly, lexical competition among items produced by different speakers is found to be less than that among items by the same speaker (see Creel et al 2008). In this respect, VFM hypothesises that alternating speakers belonging to one gender group is less disruptive than alternating both voices and gender membership. Interestingly, Palmeri et al (1993: 324) asked participants to make explicit voice-recognition judgments and found “a strong tendency to classify different-voice/same-gender repetitions [of test items] as ‘same,’ especially at short lags”.

By the same token, variability in speaking rate seems to undermine speech recognition, especially in the presence of noise. Numerous experiments demonstrate that speech recognition and recall accuracy are worse in mixed-rate conditions than in single-rate conditions (e.g., Bradlow et al 1999; Nygaard et al 1995; Sommers et al 1994). Conversely, variability in speaking rate is found to facilitate learning (e.g., Sommers & Barcroft 2007). Importantly, Miller and Volaitis (1989) and Wayland et al (1994) show that the processing of phonetic categories does not occur independently of rate information, and that changes in speaking rate effect a shift in the location of the boundary between phonetic categories and of the perceived best exemplar. As I have explained above, according to VFM, changes in speaking rate cause a phone-field to be shifted across different background positions. This could be responsible for the reduced recognition accuracy in mixed-rate conditions.

6.4 Conclusion

In this chapter, I proposed a novel approach to phonetic variability, the Variability Field Model (VFM). According to VFM, phonetic data are essentially a blend of variability sets, most exhibiting multiple sources of variation at play. To estimate variability, we need to decompose these multi-source variability sets until we reach a stage where variability sets are not further reducible to smaller component sets traceable to known sources.

I have shown how VFM gives us a new perspective on the processing and representation of variability. I focused specifically on two types of variation: allophonic and indexical variations. I discussed a VFM scheme whereby allophonic variations form context-bound phone-fields, while indexical variations provide some form of background against which a phone-field is accessed. According to VFM, phonetic data are examined as intervals rather than as single data points. In the next chapter, I propose an algorithm that allows us to examine phonetic data as intervals.

7 Variability Fields and Vowel/Zero Neutralisation in BHA

7.1 Introduction

In chapter five, I introduced and justified alternative measures of central tendency and dispersion. I argued that these measures are intuitive, cognitively more plausible, and statistically more robust than the common parametric measures that are currently used to quantify phonetic phenomena. Specifically, I suggested that the mode, rather than the arithmetic mean, be used to measure central tendency, and that the Variability Field Index (VFI), which relates the frequency of the modal interval to the range, be used to measure variability. I also argued that since the frequency of the modal interval and the number of intervals within the range are both count data, they are more in line with frequency-based Bayesian reasoning (Gigerenzer & Hoffrage 1995). They also reflect better the intuitive notions of average and variation. See chapter five for more details.

In chapter six, I provided a kind of theoretical framework for these quantitative measures. The general approach of the Variability Field Model (VFM) introduces a novel phonetics-specific quantification of phonetic data and variability. Of relevance here is the claim that a phone in its immediate phonetic context forms a variability phone-field, which has a measure of location and a measure of dispersion. These are, respectively, the modal interval, which stands for the field's gravitation, and the range of intervals, which represents the field's spread.

The current chapter is basically a preliminary implementation of the ideas outlined in chapters five and six. Specifically, the chapter (1) provides technical details of how to find the mode for phonetic data and how to calculate VFI and (2) offers a

preliminary VFM-based analysis of a portion of the neutralisation data which I analysed in terms of NHST in chapter four.

One of the main arguments emerging from this thesis is that an adequate quantification of phonetic data is one where they are examined as intervals rather than single points. In keeping with this view, I propose a binning algorithm that appeals to the familiar psychophysical notion of just noticeable difference (jnd). In order to decide on jnd ratios for the acoustic parameters investigated in this thesis, I make use of the jnd figures that are reported in various behavioural and neural studies in the literature.

Applying the binning algorithm to a carefully selected set of the neutralisation data from BHA, I run a VFM analysis and use the results to (1) re-construct a picture of the phonetics of vowel/zero neutralisation in BHA and (2) draw a comparison between VFM- and NHST-based analyses of the data. I revisit some of the issues discussed in light of the NHST results in chapter four, focusing on the apparent pattern of dis-correlation between the phonetics and phonology of neutralisation that the NHST results suggest. Recall that the NHST analysis reported in chapter four found phonologically complete neutralisation to be phonetically incomplete but phonologically incomplete neutralisation to be phonetically complete.

In contrast, the VFM-based analysis I report in this chapter finds no basis for the above-mentioned pattern of dis-correlation. According to the VFM results, the vowel/zero neutralisation through [a]-epenthesis, which is phonologically complete, is also phonetically complete. Conversely, the distinction between epenthetic [i] and lexical /i/, which survives in the phonology, also survives in the phonetics. That is, there is a close correlation between the phonetics and phonology of vowel/zero neutralisation in BHA. Moreover, the acoustic difference that the VFM-based analysis finds between [i] and /i/ is along F0, with [i]-F0 being lower than /i/-F0. The difference in this direction makes sense phonologically. I discuss this in the second half of this chapter.

The rest of the chapter proceeds as follows. In §7.2.1, I present a brief overview of the jnd, evaluating, in particular, its status in phonetic research. In §7.2.2, I justify the particular jnd figures that I use for the analysis in the study. In §7.2.3, I propose

and illustrate a jnd-based algorithm for binning phonetic data. In §7.3, I attempt a VFM-based analysis of a set of neutralisation data and conclude the chapter in §7.4.

7.2 Fixing Interval Width

7.2.1 *The jnd as a Criterion: Overview*

In psychophysics, two stimuli are said to be discriminable if their difference exceeds a differential threshold defined along some psychometric function. Loosely speaking, the smallest difference detectable at a standard probability of 75% is called a just noticeable difference (jnd). Since its introduction in psychophysical research, the notion of the jnd has evoked considerable scepticism. For example, the jnd has been found to be variable and sensitive to experimental procedures and subject-internal criteria (see e.g., Gigerenzer & Murray 1987; Kewley-Port 2001; Kewley-Port & Zheng 1999; Lehiste 1970; Rosenblith & Stevens 1953; Thurstone 1927).

Interestingly, this unreliability has also been long attributed to brain variability. For example, Solomons (1900: 234), appealing to what he describes as “the well known fact of variability of brain activity under identical stimuli”, hypothesises that a stimulus difference must exceed a threshold of brain variability to be just-noticeable. Much more recently, brain research has demonstrated that there is indeed internal brain variability and that this variability is actually beneficial (e.g., Basalyga & Salinas 2006; Boly et al 2007; McIntosh et al 2008; Stein et al 2005). See chapter six for more on this.

However, whether it is a brain variability range or a sensory range, the basic principle that a certain amount of variation goes undetected, which the jnd embodies, remains valid. Perhaps this partly explains why, in psychophysics, jnd-estimation is still a lively topic attracting a lot of dedicated research. For example, a special issue of *Perception and Psychophysics* (2001) is devoted to documenting the latest advances related to the jnd. Indeed, the jnd has not completely fallen out of favour (cf. Port 1996), nor have probability distributions or signal-detection theory completely replaced it (cf. Gigerenzer & Murray 1987).

More to the point, there is actually cognitive and neural support for the notion of jnd. For example, numerous neural studies document the existence of a negative component of the auditory event-related potential (ERP) of brain response known as mismatch negativity (MMN) (Näätänen et al 1978). MMN is elicited by any discriminable changes in auditory stimulation even in the absence of attention on the part of the subject (see Näätänen 2001, Näätänen et al 2007; Picton et al 2001 for reviews). It is these features of MMN that have led some researchers to promote MMN as an index of change detection (e.g., Näätänen & Alho 1995; Näätänen & Winkler 1999; Ylinen 2006). In this respect, the notion of representational width that is proposed by Näätänen and Alho (1995) to quantify the range of detectable auditory changes comes as a neural equivalent of the psychophysical notion of jnd. In fact, Näätänen and Alho (1995: 322) suggest that a jnd “might well correspond to a just noticeable MMN” (see also Kraus et al 1995). Representational width is still under-developed, but the principles it stands for are exactly those the jnd represents. Empirically, there is a close correlation between neural and behavioural results involving auditory discrimination (see Amenedo & Escera 2000; Näätänen et al 1993; Sams et al 1985). I review some of these results in the next section. Importantly, within phonetics and laboratory-phonology circles, the notion of jnd does not seem to have invited as much hostility as it has during its early years in psychophysics work.

In fact, in phonetic research, there is a sizeable body of literature on the subject. For example, a very recent study appearing in *Phonetica* (2007) is wholly dedicated to investigating and establishing the jnd in speech tempo (Quené 2007). A number of recent papers and textbooks still appeal to the notion of jnd when trying to assess the practical importance of an acoustic difference (e.g., Davidson & Roon 2008; Gouskova & Hall 2009; Morrison 2008; Pycha 2006; Recasens & Espinosa 2009; Remijsen & Gilley 2008).

On the use of jnd in assessing phonetic data, consider these quotes from the phonetics and phonology literature.

Nearly-identical exemplars (differing by one jnd) will be treated as identical (Pierrehumbert 2001: 141).

[E]ven if the constraint system does not take into account the issue of noticeable differences and generates candidates

that the perceptual system cannot differentiate, the differential limens [i.e., jnd] will still play a role in perception, causing the indiscriminable candidates to be learned as one (Zhang 2007: 449).

Admittedly, there are genuine methodological concerns still hanging over the reliability of the jnd. For example, the jnd is sensitive to subject training and to phonetic context (e.g., Kewley-Port 1995; Kewley-Port & Neel 2006; Lehiste 1970; Mermelstein 1978; Sommers & Kewely-Port 1996). Moreover, results obtained for steady-state portions of a very small set of vowels in a limited set of languages do not necessarily generalise to other vowels in other languages (e.g., Harris & Umeda 1987; 't Hart 1981; Klatt 1973).

However, the strategy I adopt in this thesis is this. Rather than come up with a better measure, which is beyond the scope of this thesis, I will be selective in making use of the available jnd figures from the phonetic and neural literature. I detail this strategy next.

7.2.2 The jnd as a Criterion: Figures for Parameters

The five acoustic parameters this thesis investigates are F0, intensity, duration, and F1 and F2. As I mentioned above, there are numerous behavioural and neural studies offering jnd figures for each of these acoustic parameters. However, more often than not, the reported jnd figures do not agree even when they are expressed as ratios (see below for illustrations). Many investigators have attributed this unfortunate state of affairs to the disparate experimental conditions under which the jnd figures are obtained (e.g., Kewley-Port 2001; Kewley-Port & Neel 2006; Lapid et al 2008; Lehiste 1970). This remains a major avenue for further research.

For the present, however, selecting jnd figures from the literature on the basis of an informed and explicit set of inclusion criteria seems a realistic choice. To that end, I have surveyed the behavioural and neural literature on jnd with a view to compiling a realistic set of inclusion criteria. What I have obtained can best be described as a list of preferences, given in (51) below. Of course, the list is not exhaustive but should be sufficient for the illustrative purposes of this chapter.

- (51) Jnd figures for vowel parameters are preferably
- (a) Replicated/corroborated in other studies;
 - (b) Established for speech stimuli rather than pure tones;

- (c) Established for stimuli in consonantal context rather than in isolation;
- (d) Given as a ratio/percentage, or for which a ratio/percentage can be calculated.¹⁰⁵

I have used the list in (51) to choose among the available candidate jnd figures in the literature. More specifically, the jnd figures used for the binning algorithm in the next section are those which best satisfy the preferences in (51). These figures are summarised in Table 7-1.

Parameter	jnd	Sources	
		Behavioural	Neural
F0	4%	Isačenko & Schädlich (1970); Rossi & Chafcouloff (1972); Harris & Umeda (1987)	Tiitinen et al (1994)
Intensity	2%	Nishinuma et al (1983); Flanagan (1955a); Pols (1999)	Näätänen (1992)
Duration	20%	Klatt (1976); Henry (1948)	Amenedo & Escera (2000); Kaukoranta et al (1989)
F1	13%	Mermelstein (1978); Kewley-Port & Watson (1994)	Lang et al (1990)
F2	9%	Mermelstein (1978); Kewley-Port & Watson (1994)	Lang et al (1990)

Table 7-1: Selected jnd figures for each of the acoustic parameters of the study

Under optimal listening conditions, the jnd values for pure tones fall greatly below the figures quoted in Table 7-1 (see e.g., Harris & Umeda 1987; Kochanski 2006; Lehiste 1970). For example, F0 is reported to have a jnd of .23%—.45% (Flanagan & Saslow 1958), .25% (Klatt 1973), and .6%—1.2% (Flanagan 1957). However, in

¹⁰⁵ Expressed as a ratio, the jnd relates the smallest discriminable difference found along the acoustic parameter in question (e.g., ΔF for frequency) to a reference value (base) for which responses have already been calculated in the test (e.g., F for frequency). The jnd is said to approximate Weber's Law, with the quotient (or rate) being nearly constant and the amount of detectable difference being thus proportional to the base magnitude. That is, the larger the base value, the larger the difference needs to be to get detected. So, for frequency, $\Delta F/F$ gives a jnd rate expressible as a Weber fraction or percentage. Many researchers acknowledge this property of jnd for phonetic data (e.g., Klatt & Cooper 1975; Kewley-Port & Neel 2006). Yet the constancy of the reported Weber ratios for different reference values has not been reliably established (see e.g., Carlyon & Moore 1984). However, it should be remembered that different studies in the jnd literature have employed different methodologies and different definitions of what a jnd is. In other words, the mixed jnd results in the literature might be partly due to design differences among these studies (see e.g., Lehiste 1970).

a neural study on tones by Tiitinen et al (1994), a pitch difference of .5% failed to elicit an MMN. The smallest pitch difference that evoked an MMN in their study was 1%. The MMN it elicited, however, was very faint and short. A more distinct MMN was evoked for a pitch difference of 4% upwards. Moreover, Harris and Umeda (1987) report that the jnd of F0 in naturally produced sentences is in the region of 4%—9%. The lower-bound jnd of 4% in Harris and Umeda's (1987) study agrees with the figure that Rossi and Chafcouloff (1972) have calculated using inter-quartile deviation equations for F0. Using sentential stimuli, Isačenko and Schädlich (1970) have also obtained a jnd figure of 4% for F0. Thus, I have decided to use 4% as a jnd figure for the VFM-based analysis of F0 data in this chapter.

As to intensity, the figure of 2% is an approximation of what is reported in the literature. For example, a 70dB reference vowel whose duration is 75ms is shown in Nishinuma et al (1983) to have a jnd of 2%. The same ratio is also calculable from the data given in Pols (1999) and Flanagan (1955a). A neural study by Näätänen (1992) tested brain response for intensity changes as small as 3.7%, which elicited a small MMN. Note that this does not necessarily mean that a smaller intensity difference will not elicit an MMN; it only happens that 3.7% was the smallest difference used in the test paradigm by Näätänen (1992). It would have been a different matter had this difference failed to generate an MMN. Conversely, a larger difference will naturally evoke an MMN component, as evidenced in a neural study by Schröger and Winkler (1995) generating an MMN for an intensity difference of 5.7% and 14%.

Regarding duration, the figure of 20% appears in Klatt (1976) as a lower bound, and as a mean and median of the ratios reported in Henry (1948) for reference vowels of different durations. Other studies report much smaller ratios including 2% (Huggins 1972), 2.6% (Ruhm et al 1966), and 5% (Fujisaki et al 1975). However, as Lehiste (1970: 13) observes, these small figures “represent the limit of perceptibility under optimal conditions, whereas it appears likely that in a speech condition, the just-noticeable differences established by Henry [...] may apply”. Importantly, the figure is corroborated by neural evidence from a study by Amenedo and Escera (2000) and a study by Kaukoranta et al (1989).

As to formant frequencies (F1 and F2), I use the figures reported in Mermelstein (1978), which have been calculated for reference vowels in consonantal context using data from untrained subjects. For F1, the figure I use (13%) is the mean of the figures reported for F1=300Hz and F1=600Hz. For F2, the figure I use (9%) is the mean of the figures reported for F2=2100Hz and F2=1780Hz. Kewley-Port and Watson (1994) report similar figures for untrained subjects.

Needless to say, these jnd figures are higher than those reported for synthetic vowels in isolation as in Flanagan (1955b), for instance. However, many researchers note that the jnd is larger in continuous speech than in steady-state vowels (e.g., Ghitza & Goldstein 1983; Sommers & Kewely-Port 1996). Also, Kewely-Port (1995), citing Moore (1973), observes that jnds for vowel frequencies are greater by a magnitude of 10 than those for pure tones. In neural research, a frequency change as small as 16Hz in a 1000Hz tone has been shown to elicit MMN (Sams et al 1985). Similarly, in a study by Lang et al (1990), a frequency change of 2.7% elicited MMN for good discriminators, whereas poor discriminators needed a larger magnitude of change (7.5%) to detect the difference. For some subjects, MMN was elicited for a frequency change of no less than 14% (see also Huotilainen et al 1993).

7.2.3 A jnd-Based Binning Algorithm

To divide a phonetic dataset into jnd-based intervals, we need to determine (1) a bin width (Bw) in the same units of measurement as the acoustic parameter under analysis and (2) a maximum (MMx) and a minimum (MMn) value whose paired difference (range) is divisible by Bw with a positive integer quotient and no remainder. That quotient will be the number of bins (Bn), which is a count number, for the dataset in question.

Accordingly, the Bw of a dataset along an acoustic parameter (ap) is basically the just-noticeable difference that is represented by the relevant jnd percentage (Table 7-1) of the actual base value as found for the relevant acoustic parameter. For the purposes of this thesis, the base value for each of the five acoustic parameters investigated is a Tukey's tri-mean (i.e., a median-weighted average; see below for the relevant equation) calculated over a set of 360 data points. I have chosen this parametric measure in particular because it is said to be a statistically robust measure of location. It is also more representative than the median and less

sensitive to extreme values than the mean (Tukey 1977; Rosenberger & Gasko 1983; Weisberg 1992). Furthermore, Tukey's tri-mean is easy to calculate (Hoaglin 1983).

The VFM binning algorithm in (52) applies to phone-fields as defined in the previous chapter. The algorithm should preferably be run on as large a dataset representing a phone-field as possible. In this thesis, for each phone-field examined along each of the five acoustic parameters of the study, I use a set of 360 data points to derive the relevant VFM-based statistics (see below for more).

(52) VFM's jnd-based binning algorithm

1. Determine bin width (Bw) as follows:

$$Bw_{ap} = (jnd_{ap}/100) * TM_{ap} \text{ where}$$

Bw_{ap} is the calculated bin width for a set of data along acoustic parameter ap .

jnd_{ap} is the jnd percentage for ap as given in Table 7-1.

TM_{ap} is Tukey's tri-mean for a set of data along ap calculated according to this equation:

$$TM_{ap} = \frac{1}{4}(Q_1 + 2Q_2 + Q_3)$$

Q_1 is the lower quartile; Q_2 is the median; Q_3 is the upper quartile.

2. Determine bin number (Bn) as follows:

$$Bn_{ap} = R_{ap}/Bw_{ap}, Bn \in \{1, 2, 3, 4, \dots, n\}$$

$$R_{ap} = Maximum_{ap} - Minimum_{ap}$$

3. Determine (MMx) and (MMn) as follows:

$$MMx_{ap} = Maximum_{ap} + \frac{1}{2}((Bn_{ap} * Bw_{ap}) - R_{ap})$$

$$MMn_{ap} = Minimum_{ap} - \frac{1}{2}((Bn_{ap} * Bw_{ap}) - R_{ap})$$

4. Plot the relevant dataset as a frequency histogram using the calculated figures of (Bw), (MMx) and (MMn) as settings. The modal interval is the highest bar in the histogram.
5. Find for the modal interval its midpoint (\tilde{m}) and frequency (Fm).

As an illustration of the algorithm in (52), consider the real-data summary statistics in Table 7-2 for a set of the BHA neutralisation data along *i*-duration (see the next section for details).

Acoustic parameter: <i>i</i> -duration	
Q1	63
Q2	72.8
Q3	80.5
Maximum	122
Minimum	35

Table 7-2: Summary statistics of *i*-duration data ($n=360$) including median (Q_2), upper and lower quartiles, and maximum and minimum values

For better visualisation, I have plotted the *i*-duration data in Figure 7-1 into two histograms. In histogram (i), I use SPSS default settings for bin width and maximum and minimum values. In histogram (ii), I use the settings derived from the algorithm in (52) which appear in Table 7-3 below.

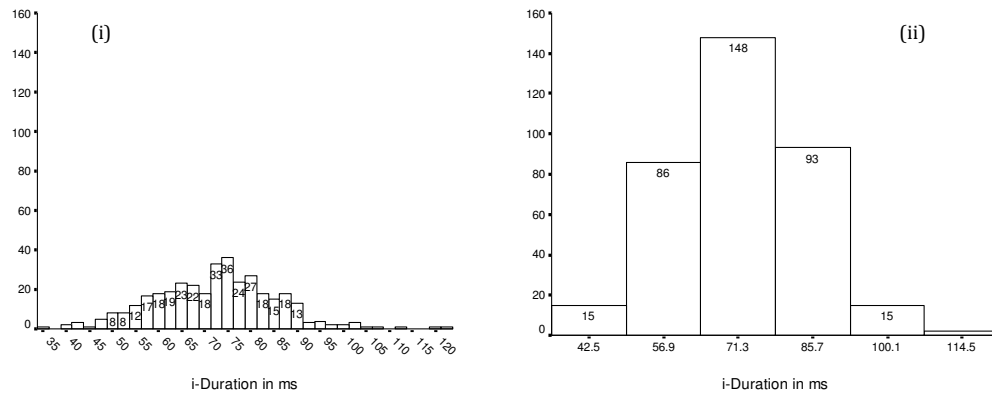


Figure 7-1: Histograms of *i*-duration data ($n=360$) generated using SPSS default settings (histogram i) and VFM settings (histogram ii); frequency scale is the same in both histograms.

Compare the figures that SPSS generates by default to the ones we get by applying the algorithm in (52) to the summary statistics in Table 7-2.

Acoustic parameter: i-duration		
	SPSS: default settings	VFM jnd-based algorithm
$Bw_{i\text{-duration}}$	2.5	14.4
$Bn_{i\text{-duration}}$	36	6
$MMx_{i\text{-duration}}$	123.75	121.7
$MMn_{i\text{-duration}}$	33.75	35.3

Table 7-3: Bin width, bin number, and maximum and minimum of i-duration data read off histograms generated using SPSS settings and VFM's jnd-based algorithm

From jnd-binned histograms, we obtain for the modal interval its midpoint (\tilde{m}) and frequency (Fm). The latter will be used to derive the Variability Field Index (VFI), VFM's measure of variability. Next, I present a preliminary VFM analysis of a set of neutralisation data from BHA.

7.3 VFM-Based Analysis of Vowel/Zero Neutralisation in BHA

7.3.1 Purpose

The main question for this VFM-based analysis is whether or not vowel/zero neutralisation in BHA is phonetically complete. In fact, this is one of the major research questions the thesis addresses. It has been termed the completeness question throughout. The analysis will also look into the nature and extent of the effect of stimulus and task manipulations on the phonetics of vowel/zero neutralisation. Results of this part of the analysis will have implications for the genuineness question, the other major research question the thesis attempts.

7.3.2 Materials

	Word	Gloss	Phone-field	Data points per condition x 2 (V ₂ -UR-Status)
1	lah/a/m	'shut tight'	aham	30x2
	lah[a]m	'meat'		
2	fah/a/m	'was out of breath'		
	fah[a]m	'char coals'		
3	nah/a/r	'yelled at'	ahar	45x2
	nah[a]r	'river'		
4	gah/a/r	'oppressed'		
	gah[a]r	'oppression'		
5	ʃah/a/r	'scalded'		
	ʃah[a]r	'month'		
6	nah/a/l	'teased'	ahal	15x2
	nah[a]l	'bees'		
7	dax/a/l	'entered'	axal	15x2
	dax[a]l	'income'		
8	rah/a/n	'pledged'	ahan	15x2
	rah[a]n	'mortgage'		
9	nah/a/r	'slaughtered'	ahar	15x2
	nah[a]r	'the act of slaughtering'		
10	ðik/i/r	'remembered'	ikir	30x2
	ðik[i]r	'prayers'		
11	fik/i/r	'came to realise'		
	fik[i]r	'thinking'		
12	gid/i/r	'managed'/'overpowered'	idir	15x2
	gid[i]r	'pot'		
13	kib/i/r	'grew'	ibir	15x2
	kib[i]r	'conceit'		
14	gab/i/l	'Gabil'	abil	15x2
	gab[i]l	'before'		

Table 7-4: Stimulus set in terms of phone-fields.

For maximum comparability with the NHST analysis presented in chapter four, I consider here the same materials used there. Recall that the NHST analysis in chapter four examined the acoustics of fourteen pairs of contrasting lexical and epenthetic vowels. Nine of these pairs illustrate neutralisation through [a]-epenthesis; the remaining five pairs illustrate neutralisation through [i]-epenthesis. Recall also that stress falls on V₁ in all the test words, which are reproduced in Table 7-4 above.

To run a VFM analysis of the acoustics of V₂ (lexical vs epenthetic) of these words, we need to consider the target vowels in terms of phone-fields rather than words. As we can see from Table 7-4, the fourteen pairs of the study comprise ten different phone-fields. To determine the modal interval of a set of data that exemplifies a phone-field, it is always preferable to consider as many data points

as possible for histogram construction (see chapter six for more). As is clear from Table 7-4, the phone-fields *ahar* and *ikir* have the largest number of data points for ‘a’ and ‘i’, respectively. Including the data collected within the six-condition paradigm in chapter four (see below for more) will bring the number of data to 540 for *ahar* and to 360 for *ikir* along each of the acoustic parameters investigated here. For comparability, the size of *ahar* data was brought down to 360 by excluding the data that come from pair number 3 in Table 7-4.

7.3.3 The Paradigm

As explained above, the dataset that I analyse in this chapter is a subset of the neutralisation data described and analysed in chapter four. Recall from chapter four that the data come from a six-condition paradigm. For the convenience of the reader, I summarise the main points here. Basically, the paradigm is a two (V_2 -UR-Status) by two (stimulus lists) by three (tasks) factorial design. Epenthetic and lexical data alike are collected in six experimental conditions manipulating stimulus materials and tasks, as illustrated in Table 7-5. There are two stimulus lists differing with respect to the order of presentation of the members of each minimal pair. These are delivered in elicitation, reading-in-context, and reading-in-a-frame tasks.

	Condition	Stimulus List	Sentential context	Task
Block I	1	Members apart	Composed sentence	Elicitation
	2	Members apart	Composed sentence	Reading
	3	Members apart	Frame sentence	Reading
Block II	4	Members in succession	Composed sentence	Elicitation
	5	Members in succession	Composed sentence	Reading
	6	Members in succession	Frame sentence	Reading

Table 7-5: Experimental conditions in blocks based on stimulus list (reproduced from chapter four)

7.3.4 Results

As we can see from the figures in Table 7-6 below, datasets along all the acoustic parameters of the study except i-F0 display uni-modality. That is, only i-F0 data have two modal intervals with an 8.6Hz difference (one jnd) separating their midpoints. This is shown in Figure 7-2. A subsequent examination of i-F0 by the underlying status of V_2 (where epenthetic [i] and lexical /i/ datasets have 180 data points each) confirms this bimodality: [i]-F0: \bar{m} =205.8Hz; /i/-F0: \bar{m} =214.4Hz. Note that F0 of lexical /i/, which is stressable in BHA, is higher than F0 of epenthetic [i],

which is not stressable in this dialect. I will comment on the directionality of this difference in the discussion section. It is important to remember that neither vowel is actually stressed in the minimal-pair stimulus set used for analysis.

	'a'		'i'	
	\tilde{m}	VFI	\tilde{m}	VFI
F0 (in Hz)	212.5	.18	214.4 205.8	.07
Intensity (in dB)	66.3	.3	61.9	.14
Duration (in ms)	82.1	.03	71.3	.03
F1 (in Hz)	833.5	.026	546	.019
F2 (in Hz)	1693	.045	2454	.026

\tilde{m} =midpoint of modal interval

Table 7-6: VFM summary statistics of the BHA neutralisation data along the acoustic parameters of the study

In addition to this central-tendency difference, epenthetic [i] and lexical /i/ have different variability magnitudes. This variability difference, too, is along F0. Specifically, the VFM analysis found [i]-F0 to be less variable than /i/-F0. This is shown in Figure 7-3. For better visualisation, I have also graphed the variability differences proportionally in Figure 7-4. As we can see, the VFI values for epenthetic and lexical vowels along most parameters are of the same (or very similar) magnitude. Most meet at the 50% line in the figure. Notable exceptions are i-F0 and a-intensity, with [i]-F0 being less variable than /i/-F0. The opposite pattern holds for a-intensity.

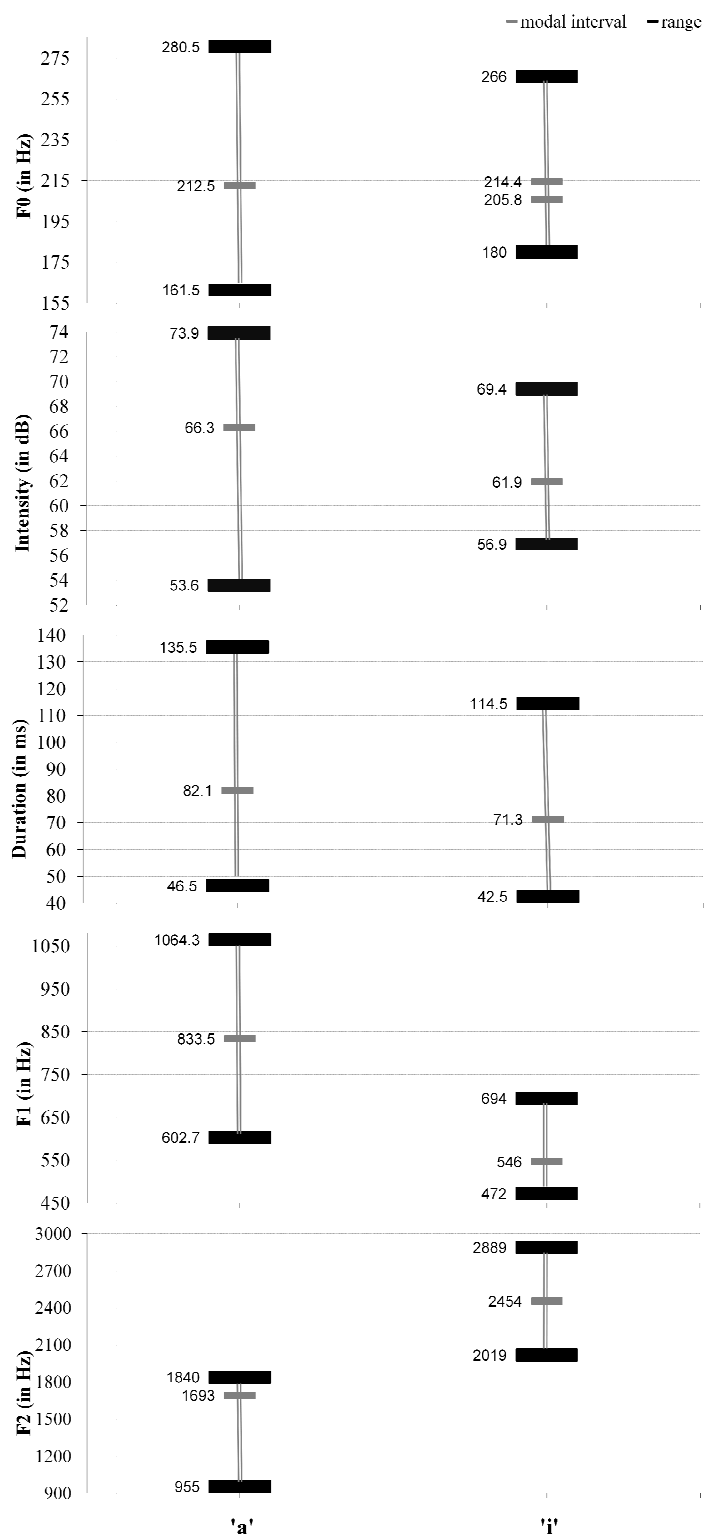


Figure 7-2: Line charts displaying modality (grey horizontal bars) and range (black horizontal bars) of the BHA neutralisation data along the five acoustic parameters of the study

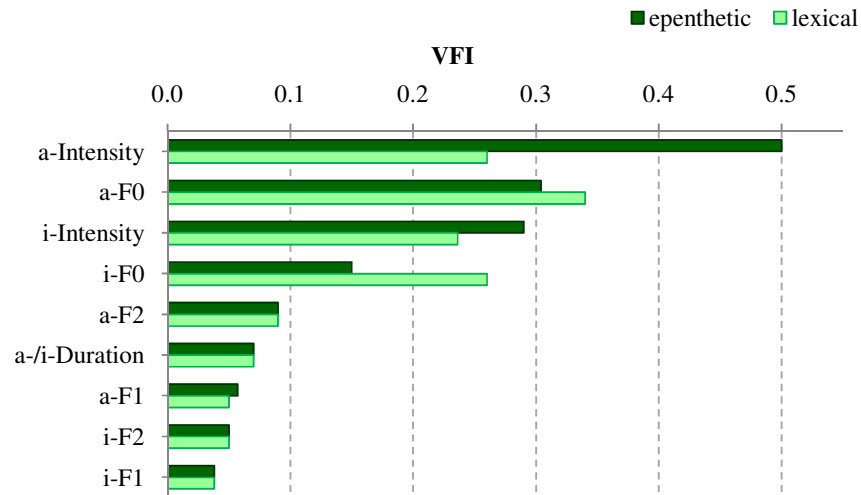


Figure 7-3: Bar chart displaying VFI values of the five acoustic parameters of the study for both 'a' and 'i' data by V_2 Underlying Status

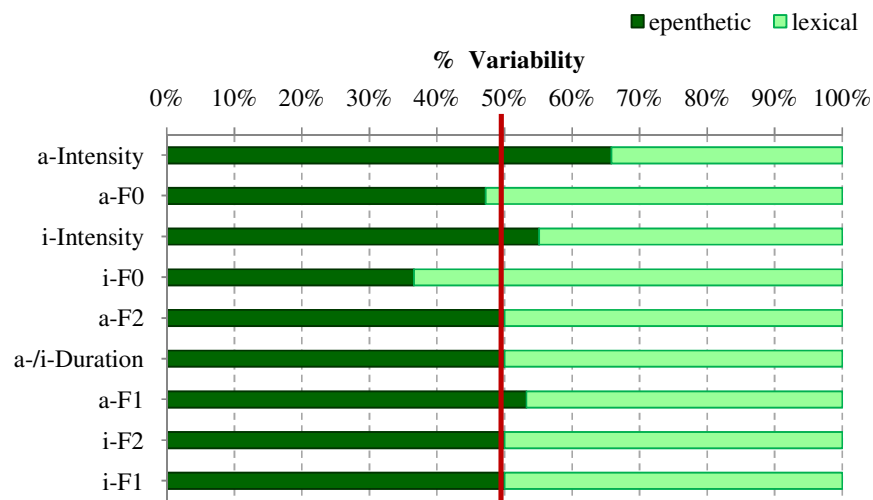


Figure 7-4: Bar chart displaying VFI values as percentages for both 'a' and 'i' data by V_2 Underlying Status.

In fact, a-intensity comes out as the most variable parameter in the study. Next comes a-F0. The least variable parameters are F1, F2, and duration for both vowels almost alike, with i-F1 showing the smallest amount of variation among all the parameters investigated. All of these effects are graphed in Figure 7-5, while proportional values are shown in Figure 7-6.

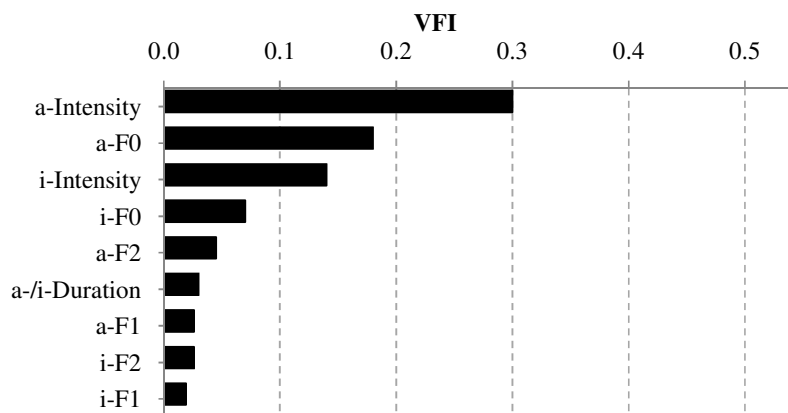


Figure 7-5: Bar chart displaying VFI values of the five acoustic parameters of the study for both 'a' and 'i' data across V_2 Underlying Status

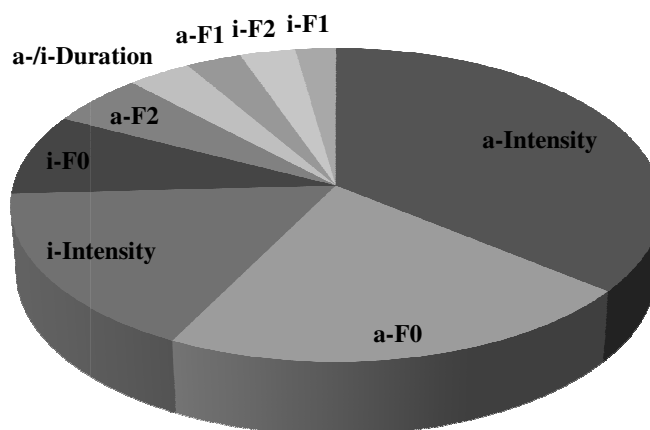


Figure 7-6: Pie chart displaying proportions of VFI values of the five acoustic parameters of the study for both 'a' and 'i' data across V_2 Underlying Status

Looking at the data across experimental conditions, we again see that the only epenthetic-lexical difference that persists across all six conditions is an F0 difference for 'i'. This is shown in Figure 7-7 through Figure 7-11. The modal-interval midpoints of [i]-F0 and /i/-F0 differ by at least 1 jnd (sometimes by 2 jnds or even 3 jnds) in all six conditions. We do not observe this pattern for any of the other parameters for either 'a' or 'i'. These results are more clearly shown in Figure 7-12. As we can see from the graphs, i-intensity seems to be the second most

epenthetic-lexical differentiating parameter. [i]-intensity and /i/-intensity are at least 1 jnd apart in all conditions except conditions 1 and 2, where bi-modality delivers a zero difference for some speakers and an 8.6Hz (1 jnd) difference for others.

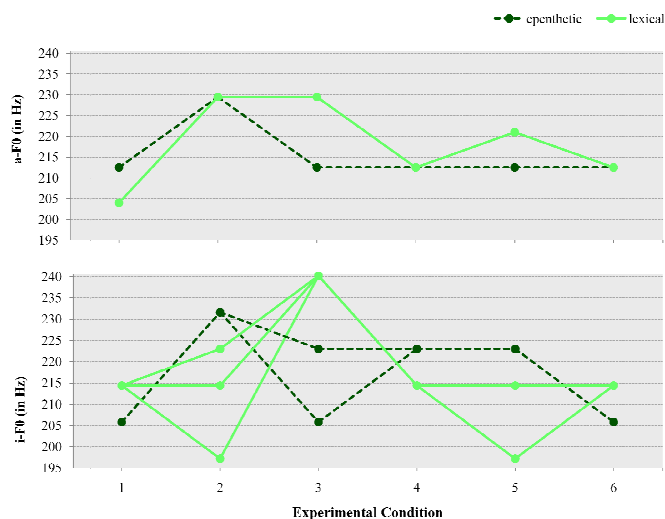


Figure 7-7: Midpoint of modal intervals of a-F0 and i-F0 data according to V_2 Underlying Status in the six experimental conditions of the study

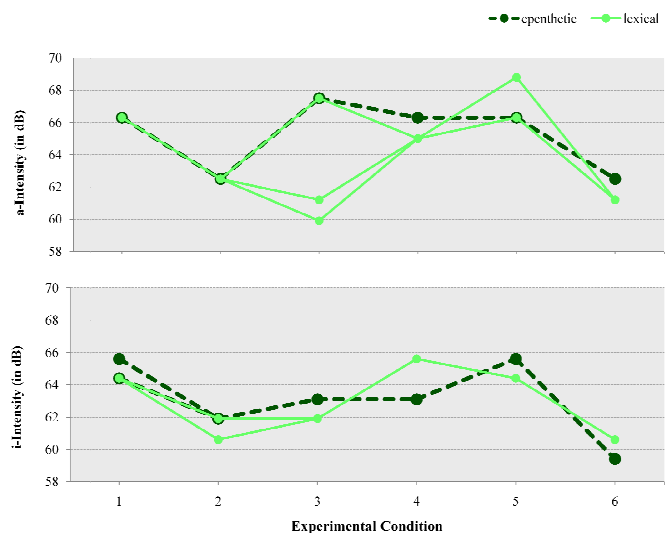


Figure 7-8: Midpoint of modal intervals of a-intensity and i-intensity data according to V_2 Underlying Status in the six experimental conditions of the study

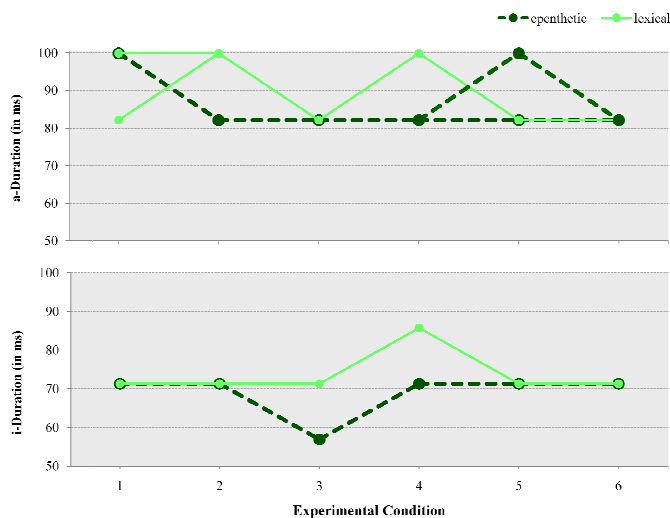


Figure 7-9: Midpoint of modal intervals of a-duration and i-duration data according to V_2 Underlying Status in the six experimental conditions of the study

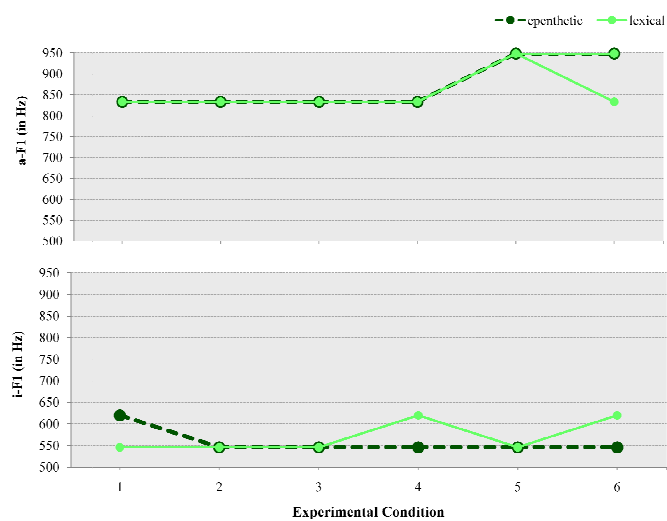


Figure 7-10: Midpoint of modal intervals of a-F1 and i-F1 data according to V_2 Underlying Status in the six experimental conditions of the study

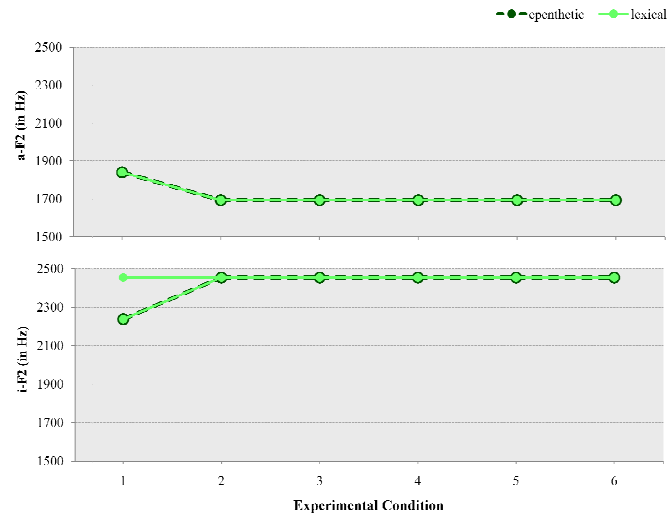


Figure 7-11: Midpoint of modal intervals of a-F2 and i-F2 data according to V_2 Underlying Status in the six experimental conditions of the study

To examine the effect of stimulus and task manipulations, we need to look at how much the modal interval of each experimental condition departs from the modal interval of the relevant phone-field across all the experimental conditions as given in Table 7-6 and Figure 7-2. These modal departures are graphed in Figure 7-13 through Figure 7-16. A departure from the modal interval of the phone-field represents the extent and direction of the effect of the experimental manipulation on the acoustics of the relevant parameter. A zero difference, on the other hand, represents resistance to experimental manipulations.

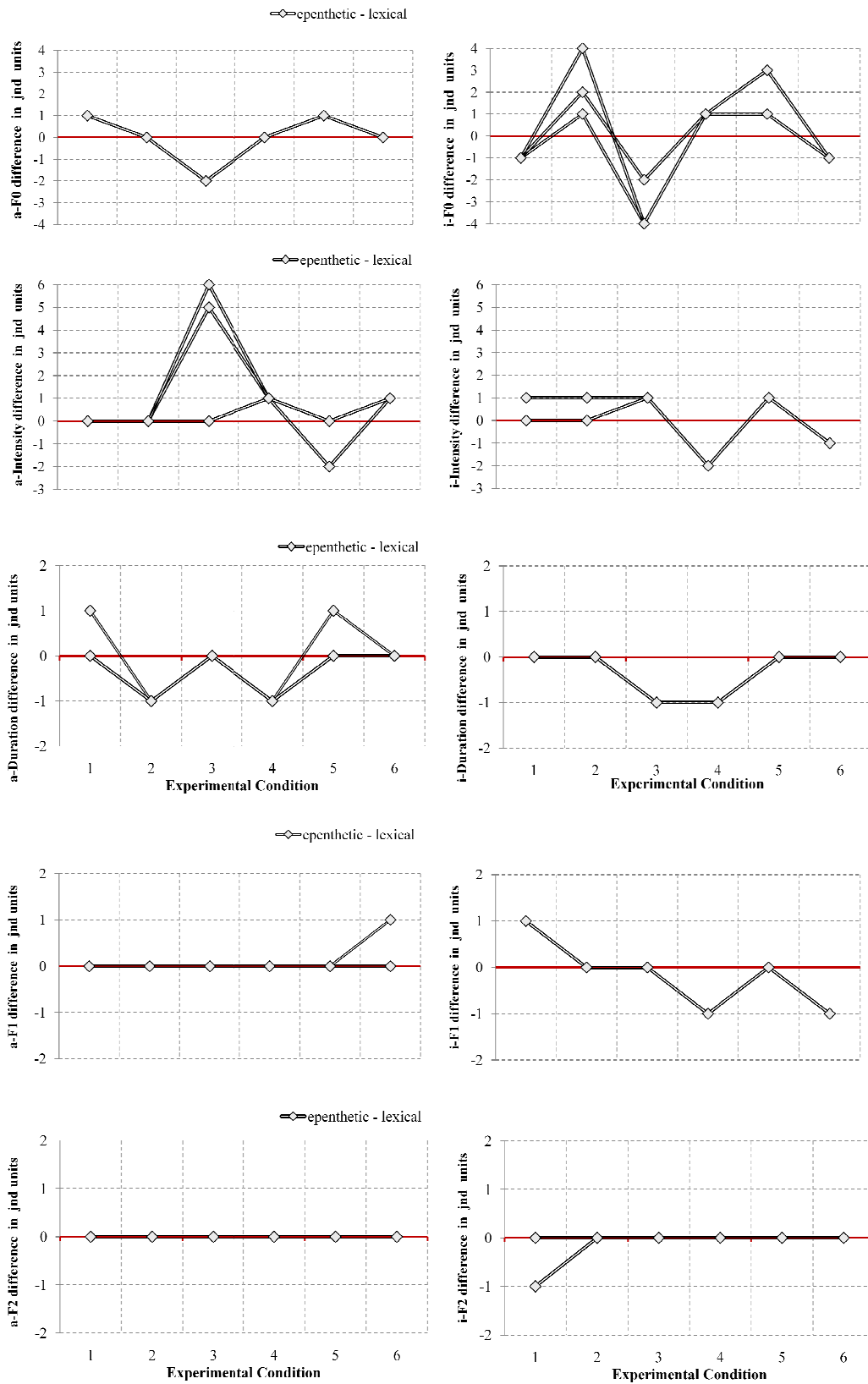


Figure 7-12: Modal differences in jnd units between epenthetic and lexical vowels in the BHA neutralisation data in the six experimental conditions of the study

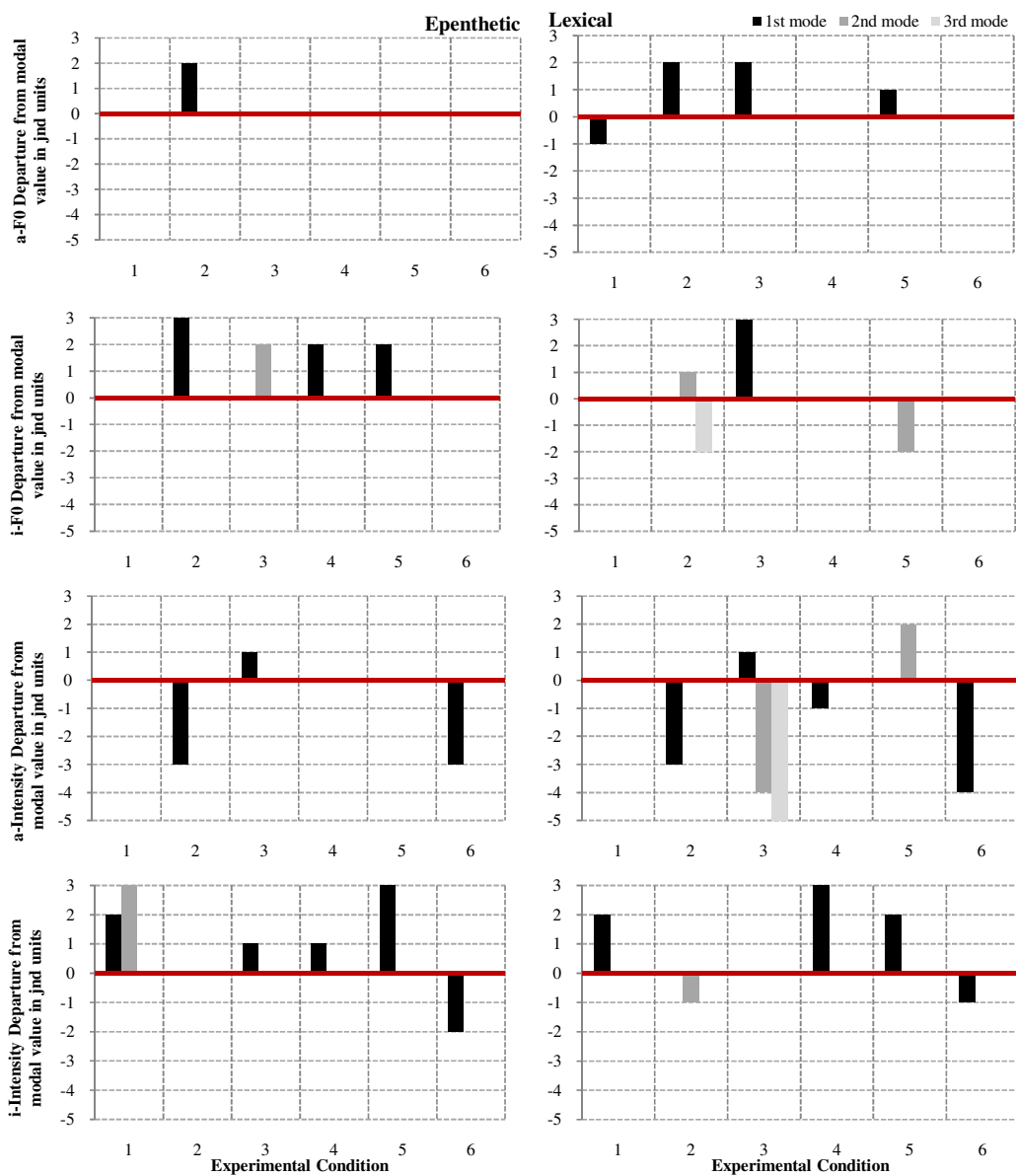


Figure 7-13: F0 and intensity modal-departures from the modal interval of the phone-field by epenthetic and lexical vowels in the six experimental conditions of the study; zero-difference lines are shown as horizontal continuous lines.

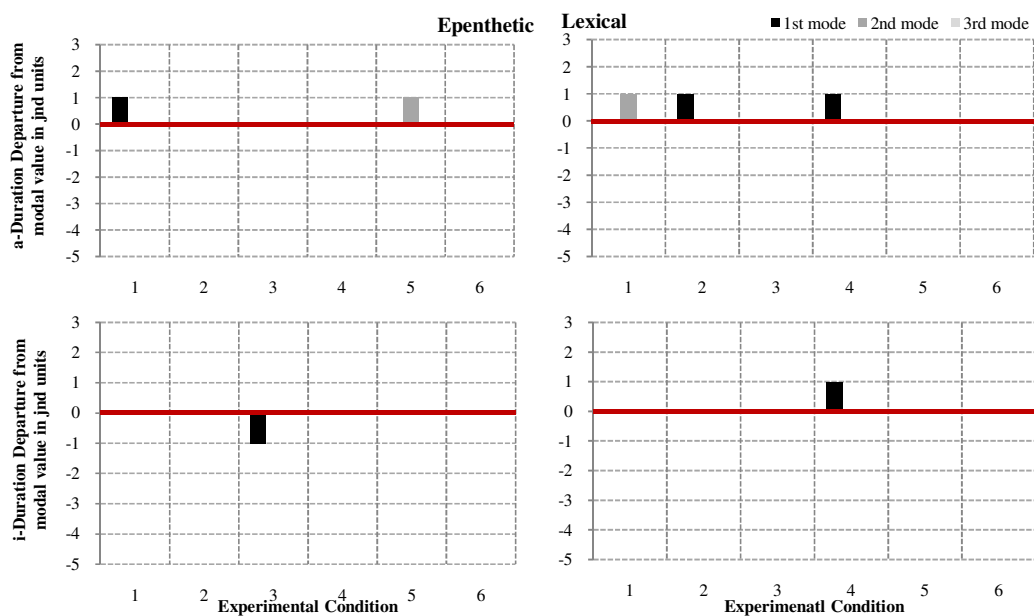


Figure 7-14: Duration modal-departures from the modal interval of the phone-field by epenthetic and lexical vowels in the six experimental conditions of the study; zero-difference lines are shown as horizontal continuous lines.

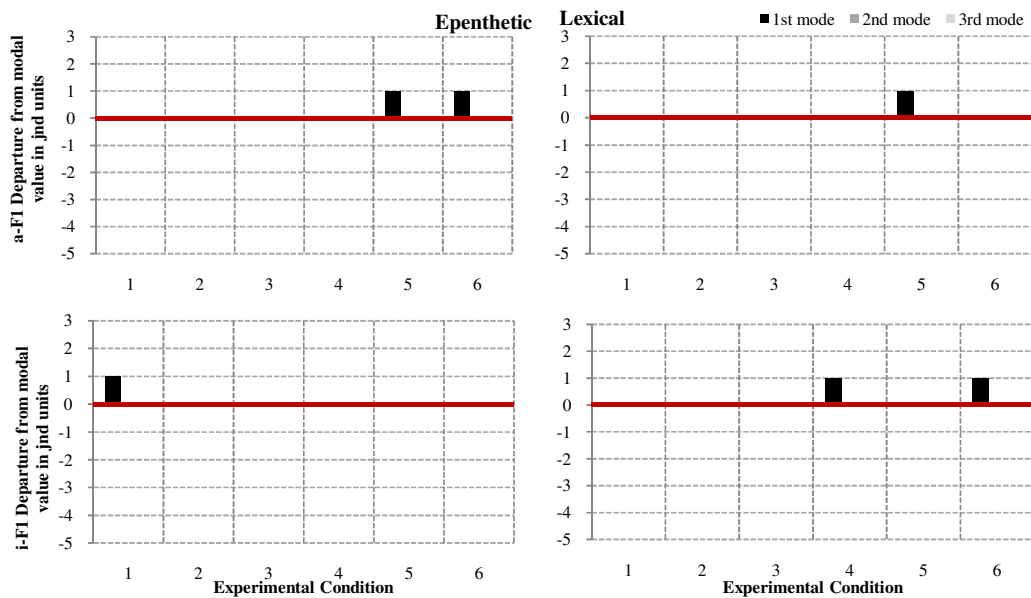


Figure 7-15: F1 modal departures from the modal interval of the phone-field by epenthetic and lexical vowels in the six experimental conditions of the study; zero-difference lines are shown as horizontal continuous lines.

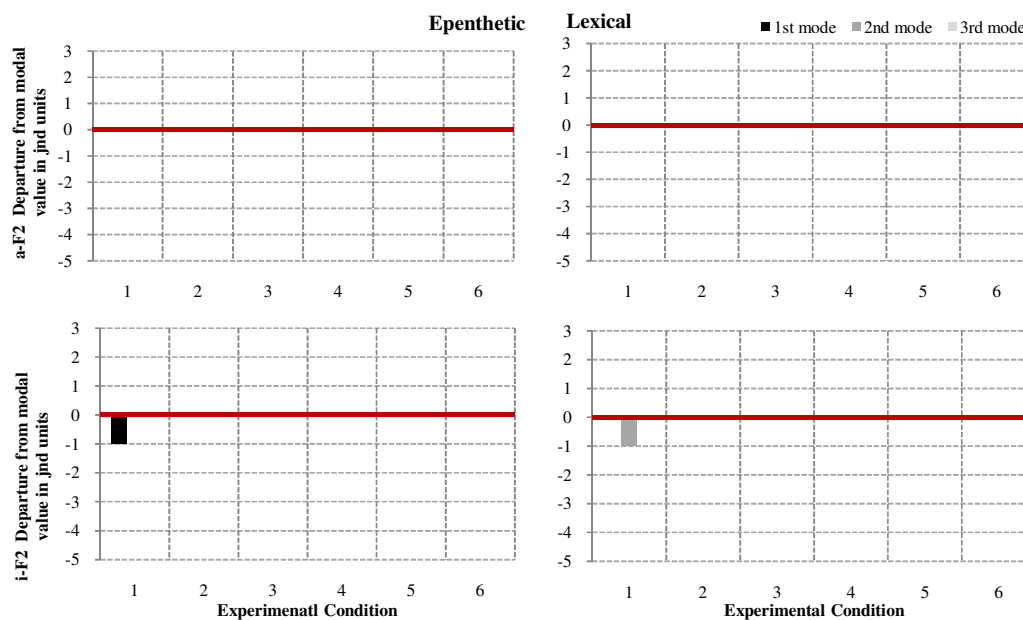


Figure 7-16: F2 modal-departures from the modal interval of the phone-field by epenthetic and lexical vowels in the six experimental conditions of the study; zero-difference lines are shown as horizontal continuous lines.

Contrasting conditions 1, 2, and 3, where members of each minimal pair in the stimulus list appear as far apart as possible, with conditions 4, 5, and 6, where members of each minimal pair appear in succession, we observe no obvious effect that can uniquely be attributed to stimulus manipulation on the modal intervals of any of the five acoustic parameters investigated here for either ‘a’ or ‘i’.

Examining task effects, we observe that /i/-F0 is lower by 2 jnds in the reading-in-context tasks (conditions 2 and 5). This difference seems to be uniquely due to the disparate sentential frames that contain the target words.

Orthography manipulation, which can be investigated by contrasting elicitation tasks with reading tasks, does not seem to have induced any influence on the phonetics of the vowel/zero neutralisation data analysed here.

With respect to variability, intensity and F0 are the most responsive to experimental manipulations, whereas F1 emerges as the least affected. I discuss these results next.

7.3.5 Discussion

In this section, I use the VFM-generated data to (1) re-construct a picture of the phonetics of vowel/zero neutralisation in BHA, (2) discuss the various implications of the results for the relation between the phonetics and phonology of neutralisation, (3) draw an empirically-informed comparison between the NHST-based and the VFM-based accounts of vowel/zero neutralisation in BHA, and (4) highlight the potential role that phonetic variability plays in phonology.

For the purposes of this chapter, neutralisation can be deemed phonetically complete if the analysed data along each of the five acoustic parameters studied here (F0, intensity, duration, F1, and F2) display uni-modality when plotted in a jnd-binned frequency histogram. With this in mind, we may conclude that the vowel/zero contrast is completely neutralised through [a]-epenthesis but not through [i]-epenthesis. That is, [a] and /a/ are acoustically indistinguishable along the parameters of the study, while there is a difference in F0 between [i] and /i/.

The conclusion we get when we abstract away from the identity of the epenthesised vowel is that vowel/zero neutralisation in BHA is acoustically both complete and incomplete. We have already reached this conclusion on the basis of the NHST analysis reported in chapter four. However, there is an important difference to be noted here.

The observed F0 difference between [i] and /i/ as revealed by the VFM analysis has a phonological reflex in BHA. As documented in chapter three, [i] and /i/ differ phonologically with respect to their capacity to bear word-stress. Specifically, unlike /i/, [i] repels stress to the extent that it is neither stressable nor metrifiable. I argued in chapter three that [i] is phonologically invisible to stress assignment in BHA. We know from the literature (see chapter four) that F0 is a major correlate of stress in Arabic (e.g., de Jong & Zawaydeh 1999; Zuraiq & Sereno 2007). F0 is also said to be the only cue of word-level stress in Arabic (Bouchhioua 2009). Here as revealed by the VFM analysis, it is F0 that seems to keep [i] and /i/ phonetically apart. The F0 difference involves both central tendency (modal values) and variability (VFI). [i]-F0 is lower than /i/-F0 by 8.6Hz, which represents 1 jnd here. This result actually agrees with the outcome of the perception test reported in chapter four. Recall that [i] and /i/ were discriminated more accurately than were [a] and /a/. That is, the F0 difference between [i] and /i/ is acoustically and

perceptually real. It is, nonetheless, just-perceptible. So, calling the effect incomplete neutralisation seems appropriate. Importantly, epenthetic [i] comes out as having the lowest pitch, while lexical /i/ the highest, with epenthetic [a] and lexical /a/ equally in between.¹⁰⁶ As documented in chapter three, of these vowels, only [i] is neither stressable nor metrifiable.

There is empirical evidence that a high F0 correlates with stressedness. For example, Topintzi (2006) presents experimental data where vowels following a voiceless C are perceived as stressed far more often than vowels following a voiced C, with both in exactly the same prosodic environment. She concludes that “the pitch raising caused to the vowel after a voiceless onset can be interpreted as stress” (p. 314). Also, Fry (1958) reports a similar effect for synthesised stimuli: increasing F0 results in the vowel being likely perceived as stressed (see also Morton & Jassem 1965).

Of particular significance in this regard is the association between high F0 and stress in Arabic that de Jong and Zawaydeh (1999) found. Specifically, they report that stressed vowels in Arabic have a higher F0 than the corresponding unstressed vowels. The interaction of stress and tone in the phonology of languages as diverse as Ayutle, Golin, Lithuanian, Serbo-Croatian, Vedic Sanskrit, and Hixkaryana (de Lacy 2002a, 2002b; see also Yip 2002) further supports the pattern of correlation between high F0 and stressability as documented above. Generally, in these languages, a vowel bearing a high tone (H) attracts stress.

Note that relying on the NHST-generated results of the neutralisation data in BHA, as we did in chapter four, would not allow us to locate any kind of phonetic expression for the phonological invisibility of epenthetic [i] as opposed to lexical /i/, epenthetic [a], and lexical /a/. Instead, the NHST outcome would have left us puzzling over what appeared to be a counter-intuitive pattern of anti-correlation between the phonetics and phonology of neutralisation. Recall that the outcome of NHST suggests that the vowel/zero neutralisation through [a]-epenthesis, which is

¹⁰⁶ Of relevance here is what is known in the literature as intrinsic fundamental frequency (e.g., Fischer-Jørgensen 1990; Whalen & Levitt 1995). All other things being equal, high vowels are said to have a higher intrinsic F0 than low vowels. Although the results we have for lexical /i/ and both epenthetic [a] and lexical /a/ are consistent with this observation, we know that, in this study, not all other things are equal for high and low vowels. For example, ‘i’ and ‘a’ data have not been extracted from exactly the same segmental environments. Note that this does not invalidate the results for epenthetic [i] as opposed to lexical /i/, both of which occur in exactly the same segmental environment. See also my refutation of the morpho-syntactic argument in chapter four.

phonologically complete, is acoustically incomplete. Conversely, the distinction between epenthetic [i] and lexical /i/, which survives in the phonology, is lost in the acoustic signal. That is, the phonetics and phonology of vowel/zero neutralisation in BHA dis-correlate for both 'a' and 'i'.

More puzzlingly, the acoustic parameter that NHST has pronounced as the seat of the distinction that remains of the phonological contrast between epenthetic [a] and lexical /a/ is contrastively inert. With particular reference to stress correlates, intensity is secondary to F0 in many languages (e.g., Lehiste 1970 for a review; but see Kochansky et al 2005 for a different view).

Note also that even Cohen's *d* figures (given in chapter four) appear to point towards the same conclusion as inferential statistics. Recall that a-intensity has by far the largest *d* value ($d=.811$). The next largest was i-intensity ($d=.46$). This is partly due to the small variation in intensity as measured in terms of the standard deviation (SD) and coefficient of variance (C_v).

But as we have seen in the current chapter, intensity is by far the most variable parameter. Not only does intensity score the largest VFI value, it is also the most susceptible to the various experimental manipulations in the six-condition paradigm of the study. Vowel intensity varies from condition to condition. The next most variable parameter in terms of both VFI figures and experimental fluctuations is F0. In contrast, duration and F1 and F2 come out as the least variable, again in terms of VFI figures and resistance to experimental manipulations. Actually, the VFM-generated results seem to suggest that the acoustic parameters along which many languages define their systems of contrast are precisely those which show very little variability.¹⁰⁷ Quantity and quality distinctions in vowel systems in many languages reside in these parameters (cf. tone languages)¹⁰⁸.

There are also finer distinctions to be drawn on the basis of phonetic variability as measured in terms of VFI. Specifically, i-F1 and i-F2 are less variable than respectively a-F1 and a-F2, with i-F1 being by far the least variable. That 'i' should

¹⁰⁷ The inspiration for this idea came from a discussion I had with Moira Yip.

¹⁰⁸ Note that between the two most variable parameters, it is F0 that is less variable. This is significant even in the context of tone languages, where F0 perturbations serve a contrastive function.

display less variation than ‘a’ is expected given the accumulating evidence from acoustic, articulatory, and perceptual studies, all highlighting the relative ‘stability’ of ‘i’ as opposed to ‘a’. For example, in a very recent cross-linguistic study by Gendrot and Adda-Decker (2007), ‘a’ is found to be acoustically “very variable”, while ‘i’ is highly stable. Similar results are obtained by Al-Tamimi and Ferragne (2005) for French and two dialects of Arabic. ‘i’ is also described as intrinsically “resistant to variability and coarticulation [...and] to effects of prosodic structure” (Tabain & Perrier 2005: 96). Moreover, Pearce (2007: 39), studying the effect of variation in duration on F1 and F2 values of vowels in Kera, observes that “high vowels [including /i/] are relatively unaffected by [variation in] duration”.

Likewise, articulatory (EMA) studies show that ‘i’ displays less variation than ‘a’ when they appear in different prosodic positions (Cho 2004; Tabain & Perrier 2005). There are claims in the literature that a ‘saturation effect’ in the production of /i/ (e.g., Perkell 1996; Tabain & Perrier 2005) is responsible for the little variability in F1.

Similarly, in a number of perception studies, /i/ is reported to be the most easily identifiable vowel from the burst of a preceding stop (Cullinan & Tekieli 1979)¹⁰⁹ or from the whispered transients of the burst of a preceding stop (Repp & Lin 1989). Winitz et al (1972) report a high rate of correct identification of ‘i’ compared to ‘a’ from the burst of voiceless stops /p, t, k/. In these studies, it could be argued that the burst spectra in, say, /ti/ and /ki/ sequences might carry important cues to the front high vowel, whereas no comparable cues are found for [a] in /ta/ and /ka/ sequences. However, Bonneau (1996: 2507), studying the identification of /i, u, a/ from the bursts of initial stops /p, t, k/ by French-speaking listeners, reports the following:

The subjects spontaneously told us that they could not actually recognize the timbre of the vowel /a/ while they sometimes clearly identified the timbre of the vowels /i/ and /u/. Paradoxically, the vowel /a/ obtained a high identification rate. In fact, it seems that listeners have chosen the /a/-response in the absence of a clear vocalic

¹⁰⁹ Abstracting away from the lax-tense distinction, which is one of the vowel variables in the Cullinan and Tekieli study, we observe that the high front vowel ‘i’ is identified most accurately. Among the lax vowels, /ɪ/ achieved the highest percentage of correct responses. Among the tense group, however, /ɑ/ is identified far better than /i/, which is repeatedly misidentified as its lax counterpart.

timbre. [...] It thus appears that the /a/-response has been given for the vowel /a/ and for the vowels /i/ and /u/ which were not clearly identifiable.

The acoustic variability of /a/ as described above seems to be reflected in its flexible perception. Listeners seem to be more willing to categorise as /a/ than to give an /i/-category in situations where they are merely guessing. This is significant given the results of the VFM analysis in the present study where 'a' phone-fields have larger spread areas than do 'i' phone-fields along most of the acoustic parameters of the study. A phone-field with a large spread area allows a larger amount of acoustic variation than does a phone-field with a small spread area. In the Bonneau's study, it could be argued that those instances of /i/ data that were not 'clearly identifiable' fell outside of the bounds of the small-area /i/-fields. They were misidentified as /a/ presumably because /a/-fields cover a lot of variation within comparatively large spread areas. We have already seen the consequences of spread differences for the perceptual discrimination of vocalic differences in the current study. Recall that epenthetic [i] and lexical /i/ were discriminated more consistently than were epenthetic [a] and lexical /a/. This interpretation does not conflict with the fact that epenthetic [i] and lexical /i/ are acoustically more different than are epenthetic [a] and lexical /a/. In this thesis, differences in central tendency and in variability both contribute to our assessment of the phonetics of neutralisation.

7.4 Conclusion

In this chapter, I presented a preliminary implementation of the VFM-based quantification of phonetic data. I provided technical details of how to find the mode for phonetic data and how to calculate VFI. I also offered a VFM analysis of a set of vowel/zero neutralisation data from BHA.

According to the VFM results, vowel/zero neutralisation through [a]-epenthesis, which is phonologically complete, is also phonetically complete. Conversely, the distinction between epenthetic [i] and lexical /i/, which survives in the phonology, also survives in the phonetics. Accordingly, we may conclude that there is a close correlation between the phonetics and phonology of vowel/zero neutralisation in BHA.

8 Summary and Conclusions

This thesis has sought a new, better-informed understanding of the phonetics of neutralisation. It has reached for insights from research fields as diverse as statistics, cognition, neurology, and psychophysics, in addition, of course, to phonetics and phonology—insights that have helped sharpen our theoretical and empirical perspectives on the phonetics of neutralisation. The conception of the ideas and arguments presented throughout the thesis owes a lot to a detailed survey of the empirical literature on the phonetics of neutralisation. These conceptualisations have actually shaped the landscape of the thesis.

One of the issues that figures prominently in the literature and which is in need of addressing is the issue of variability. Variability in the phonetics of neutralisation comes in two senses: an inherent quantitative variability and an acquired qualitative variability. The latter follows from our drawing a qualitative distinction between phonetically complete and incomplete neutralisation. To elucidate qualitative variability, the thesis presented a scrutiny of the labelling criteria upon which the complete-incomplete distinction is based.

Among the questions that the issue of variability has raised are the following. What are the reasons behind variability? How profitable is it to pursue the causes of variability? How do the various models of the phonetics of neutralisation approach the issue of variability? And how else should they approach variability? The thesis has devoted two chapters to discussing the last two of these questions.

Specifically, chapter two offered a critical evaluation of the existing approaches to qualitative variability in the literature. These approaches fall into three main groups with respect to their underlying conception of the phonetics of neutralisation. One group only recognises phonetically complete neutralisation as both genuine and relevant (e.g., Steriade 1999); another group only predicts

phonetically incomplete neutralisation (e.g., Ernestus & Baayen 2006); and a third group is a combination of the first two, with complete neutralisation accepted in certain cases, and incomplete neutralisation in others (Barnes 2006). All of these approaches display a great deal of intolerance of qualitative variability, which some suggest might be due to some 'paralinguistic contamination problem' (e.g., Barnes 2006: 225), 'spelling pronunciation' (e.g., Fourakis & Iverson 1984; Warner et al 2006), and/or 'hypercorrection' (e.g., Jassem & Richter 1989).

To complement the discussion of variability in chapter two, I focused in chapter six on variability in a quantitative sense. I sketched a new approach to phonetic variability where data have the capacity to form variability fields. Under this approach, variability is seen as the essence of phonetic data rather than some isolatable addition. On this conceptualisation, even random variations merit phonetic investigation. I also discussed allophonic and indexical variations, which are said to pose a real challenge to our understanding and modelling of phonetic data. The VFM schema I proposed allows only allophonic variations to form context-bound phone-fields; indexical variations, however, may only provide some form of background against which phone-fields are accessed. I presented both behavioural and neural data to support my claims.

In chapter five, I offered a qualitative and quantitative description of the phonetics of neutralisation. I first presented a brief scrutiny of the labelling criteria in the literature as applied to the phonetics of neutralisation. Then I discussed statistical significance, the single most important criterion that has been standardly used in the qualitative description of the phonetics of neutralisation. My main findings in this regard are that a statistically significant difference can be a size effect rather than an effect size, that statistical significance can easily be misinterpreted, and that the null hypothesis of no difference is almost always false on both logical and empirical grounds.

On the quantitative side, I evaluated the parametric measures of central tendency and dispersion that have been commonly used to quantify the phonetics of neutralisation. I showed that these measures are unintuitive and so closely tied to the numerical values of the measurement scale as to potentially undermine any robust estimation of the underlying central location and variability. I also proposed an alternative that is both more intuitive and cognitively plausible. Specifically, I

suggested that the mode, rather than the mean, be used to measure central tendency. This suggestion has rested on the claim that the mode reflects better the intuitive notion of average. By the same token, the variability measure I proposed relates the frequency of the modal interval to the range, both of which are more in line with frequency-based Bayesian reasoning (Gigerenzer & Hoffrage 1995). Moreover, I suggested that phonetic data are more appropriately examined as intervals rather than as single points. In chapter seven, I proposed a binning algorithm utilising the familiar psycho-physical notion of just noticeable difference (jnd).

The variability issue has actually guided the theoretical focus of the thesis and provided at the same time a context for the experimental part of the thesis. This has been designed to address the two main issues that have occupied the bulk of the phonetic literature on neutralisation. These issues concern the extent of phonetic merger (the completeness question) and the empirical validity of the phonetic effect (the genuineness question). Regarding the completeness question, I presented in chapter four acoustic and perceptual analyses of vowel/zero alternations in Bedouin Hijazi Arabic (BHA). As documented in chapter three, the phonology of these alternations exemplifies two neutralisation scenarios bearing on the completeness question. To the best of my knowledge, this thesis is the first to have looked closely at both scenarios together, testing hypotheses involving the acoustics-perception relation and the phonetics-phonology relation.

The NHST analysis of the production experiment reported in chapter four reveals a curious pattern of dis-correlation between the phonetics and phonology of neutralisation. There is a statistically significant acoustic difference between epenthetic [a] and lexical /a/ that the phonology seems to overlook. At the same time, the phonology treats epenthetic [i] and lexical /i/ differently despite their acoustic non-distinguishability by the criterion of statistical significance. Moreover, the acoustic difference that reaches statistical significance is an intensity difference between epenthetic [a] and lexical /a/. In other words, the only statistically significant difference is along a phonetic parameter that is contrastively inert in many languages, most certainly in BHA.

The results of the perceptual analysis suggest that listeners discriminate epenthetic [i] and lexical /i/ more accurately than they discriminate epenthetic [a]

and lexical /a/, a finding that sits uncomfortably with the picture painted by the NHST-generated inferences regarding the acoustics of epenthetic and lexical vowels in BHA. The results are not consistent with the phonological account in chapter three either.

Next, I discussed the genuineness question from an experimental and statistical perspective. Experimentally, I devised a paradigm that manipulates important variables claimed to influence the phonetics of neutralisation. These variables include stimulus composition, orthography, and pragmatic context. The main finding in this regard is that the same experimental make-up can produce both complete and incomplete effects, with no definitive correlation between either effect and experimental artefactuality.

Statistically, I re-analysed neutralisation data reported in the literature from Turkish and Polish, applying different pre-analysis procedures. My main finding here is that different statistical analyses of the same neutralisation data can yield qualitatively different results. This led me to conclude that arguments questioning the genuineness of the reported findings are self-defeating, irrespective of whether they appeal to experimental or statistical considerations.

In chapter seven, I offered a VFM analysis of a portion of the acoustic data which were NHST-analysed in chapter four. According to the VFM results, the neutralisation effect through [a]-epenthesis, which is phonologically complete, is also phonetically complete. Conversely, the distinction between epenthetic [i] and lexical /i/, which survives in the phonology, also survives in the phonetics. In other words, the VFM analysis yields a close correlation between the phonetics and phonology of vowel/zero neutralisation in BHA. Moreover, the acoustic difference that the VFM analysis found is an F0 difference between [i] and /i/, with [i]-F0 being lower than /i/-F0. This result in terms of both the dimension itself and its directionality is phonologically expected.

A key conclusion to draw from the thesis is that there is good reason to reconsider many of the questions asked in the phonetic literature on neutralisation in terms of a different perspective—one that places emphasis on variability.

A recommendation emerging from the thesis is that we need to encourage less reliance on NHST and open up to other analytical techniques, including apparently subjective and fuzzy approaches that rely on intuitions for scientific inference.

The VFM alternative sketched in this thesis is specifically tailored to phonetic data. There are at least two features that set VFM apart from competing models of lexical representation and phonetic processing. Firstly, VFM takes a middle-ground approach to phonetic variability. While VFM celebrates phonetic variability and advertises it as the essence of phonetic data, it prunes down phonetic data so that we do not end up drowning in pools of phonetic variation that includes differences that are not even just-noticeable. A central VFM argument is that to find structure in variability, we need only consider hearable variability. Secondly, for summarising a set of phonetic data in a parametric fashion, VFM places emphasis on count data rather than arithmetically derived data like the mean and SD. This manoeuvre on the part of VFM brings the model more in line with frequency-based Bayesian reasoning (Gigerenzer & Hoffrage 1995) and with intuitive conceptions of the notions of average and variability.

Some of the notions that VFM appeals to are still in need of refinement and validation. Nonetheless, the model has clear potential of opening up new avenues for further research.

An obvious direction for future research is to develop VFM components and implementation tools. An important starting point is the jnd, which plays a central role in defining an interval width according to which datasets are divided into bins, which in turn determine the data mode and spread. We still need to work further on refining jnd measures for the various acoustic dimensions of speech sounds that we study. Real improvement needs to find its way into design, methodology, and analysis. The jnd-estimation studies in our field should apply the latest methodological advances made by psychophysics and other relevant research fields (for an overview see e.g., Klein 2001; Leek 2001). As to evaluation, researchers concerned should look for replications. An important contribution to this research programme is the utilisation of behavioural and neural data collected within the same experimental paradigm for cross-validation. Here, the neural component known as MMN, which is currently advertised by many as an automatic change-detection index, comes to mind. Neural paradigms need not test arbitrary

and fixed differences between baseline and deviant stimuli (e.g., Gomes et al 1995: [100ms vs 170ms]; Huotilainen et al 1993: [a difference of 150Hz]; Inouchi et al 2003: [a difference of 152ms]; Kaukoranta et al 1989: [100ms vs 50ms]; Todd & Michie 2000: [50ms vs 125ms]; but see Näätänen 1992; Tiitinen et al 1994). Instead, differences should preferably vary across a range of values in small-step increments/decrements. Collaboration between neurologists and phoneticians/phonologists is a promising enterprise for the advancement of scientific inquiry into this issue.

Another avenue for further research, suggested by the findings of this thesis, concerns the dimensionality issue of phonetic data. This carries special importance for the question of representation and processing of phonetic data in terms of VFM. There is a sizeable body of behavioural and neural literature dealing with the dimensionality issue (e.g., Garner 1974, 1970; Garner & Felfoldy 1970; Giard et al 1995; Gomes et al 1995; Katseff & Houde 2008; Wood 1974; Turk & Sawusch 1996). A particularly important question here is how to represent data belonging to a given phone-field which have been derived from a number of parameters, such as duration, F0, and intensity. More generally, are stimulus data along different phonetic parameters processed (and thus presumably represented) integrally or separably in the sense of Garner (1974)? Is there symmetry in such processing and representation (e.g., Garner 1983; Turk & Sawusch 1996)? How does this impact on VFM?

Another issue into which more research needs to be channelled is the dynamicity of phonetic data. Many phoneticians agree that phonetic data should be examined along both dynamic and static parameters (e.g., Lindblom 1990; Meunier et al 2006). Progress has already been made in harnessing the dynamicity of phonetic variability, for example, in the study of speaker characteristics (e.g., McDougall 2006, 2004; Nolan et al 2006) and the application of nonlinear dynamical systems for the analysis of phonetic data (e.g., Gafos 2006; Gafos & Benus 2006; Nguyen et al 2009; Nycz 2005; Tuller et al 2008, 1994).

It is clearly desirable to integrate the dynamicity and dimensionality issues in our future attempts to learn more about phonetic phenomena, including the phonetics of neutralisation. The approach I advocate here examines phonetic data as intervals rather than single points, and uses the mode as a measure of central

tendency in place of the mean. This thesis has already demonstrated the profitability of this approach.

APPENDIX A

A non-exhaustive list of the phonological processes that have been investigated in the literature on the acoustics of neutralisation. The list does not include articulatory or perception studies.

Note:

Incomplete neutralisation= there are statistically significant differences between the sounds that are subject to neutralisation.

Complete neutralisation= there are no statistically significant differences between the sounds that are subject to neutralisation.

Variable= neutralisation is incomplete in certain conditions and complete in other conditions (e.g., certain segment types, experimental tasks, pragmatic contexts, etc.).

NoST= No statistical significance testing reported.

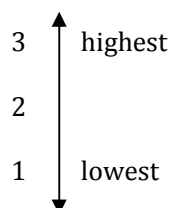
Phonological Process	Language	Study	Finding: Neutralisation is
Final Devoicing	German	Piroth & Janker 2004	Complete
		Piroth et al 1991	Variable
		Port & Crawford 1989	Incomplete
		Port & O'Dell 1985	Incomplete
		Charles-Luce 1985	Variable
		Fourakis & Iverson 1984	Variable
		Mitleb 1981	Incomplete
		Taylor 1975	NoST
	Catalan	Dinnsen & Charles-Luce 1984	Variable
		Charles-Luce & Dinnsen 1987	Incomplete
		Charles-Luce 1993	Variable
	Polish	Slowiaczek & Dinnsen 1985	Incomplete
		Tieszyn 1997	Incomplete (two out of three dialects)
		Jassem & Richter 1989	Variable
		Giannini & Cinque 1978	NoST
	Russian	Pye 1986 (in Kopkallı 1993)	NoST
		Chen 1970	NoST
		Dmitrieva et al 2010	Incomplete
	Dutch	Ernestus & Baayen 2006	Variable
		Warner et al 2004	Incomplete
Jongman 2004		Incomplete	
Baumann 1995		Complete	
Turkish	Kopkallı 1993	Complete	
Lithuanian	Campos-Astorkiza 2008	Incomplete	
Kuala Lumpur Malay	Abu Bakar et al 2007	NoST	
Friulian	Baroni & Vanelli 2000	Incomplete	
Afrikaans	van Rooy et al 2003	Incomplete	
	van Rooy & Wissing 1996 (in van Rooy et al 2003)	Variable	
vowel-deletion and vowel-reduction	Russian	Padgett & Tabain 2005	Variable
		Barnes 2006	Variable
	French	Fougeron & Steriade 1997	Incomplete
	Catalan	Herrick 2004	Complete
	Shimakonde	Liphola 2001	Complete
Serbo-Croatian	Ridjanovic 1986	NoST	

Phonological Process	Language	Study	Finding: Neutralisation is
flapping	American English	Herd et al 2010	Incomplete
		Braver 2010	Incomplete
		Sharf 1962	NoST
		Huff 1980	Variable
		Zue & Laferriere 1979	Incomplete
		Fox & Terbeek 1977	Incomplete
		Fisher & Hirsh 1976	Variable
		Port 1977	Complete
stop-epenthesis	English	Sheldon 1973	Variable
		Lee 1991	NoST
		Arvaniti 2006	Incomplete
		Arvaniti & Kilpatrick 2007	Complete
		Fourakis 1980	Incomplete
		Yoo & Blankenship 2003	Complete
assimilation	English	Fourakis & Port 1986	Incomplete
		Torres 2001	Variable
	French	Snoeren et al 2006	Incomplete
	Catalan	Charles-Luce 1993	Variable
		Torres 2001	Variable
	Russian	Burton & Robblee 1997	Incomplete
	Lithuanian	Campos-Astorkiza 2008	Incomplete
Bengali	Lahiri & Hankamer 1988	Complete	
coda-neutralisation	Eastern Andalusian Spanish	Gerfen & Hall 2001	Incomplete
	Western Andalusian Spanish	Rueda-López 2007	NoST
	Puerto Rican Spanish	Simonel et al 2008	Incomplete
	Korean	Kim & Jongman 1996	Complete
vowel epenthesis	Lebanese Arabic	Gouskova & Hall 2009	Incomplete
	Palestinian Arabic	Gouskova & Hall 2007	Complete
	Brazilian Portuguese	Gristófaro-Silva & Almeida 2008	Incomplete
consonant deletion	Turkish	Rudin 1980	Incomplete
enchaînement	French	Fougeron 2007	Incomplete
vowel length	Dutch	Lahiri et al 1987	Complete
singletons vs geminates		Warner et al 2004	Incomplete
verb allomorphy		Warner et al 2006	Complete
tone sandhi & contrasts	Mandarin	Peng 2000	Incomplete
	Cantonese	Yu 2007	Incomplete
	Taiwanese	Myers & Tsay 2008	Variable

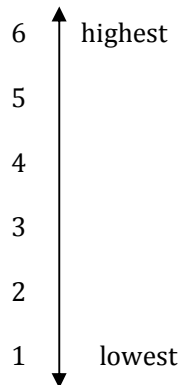
APPENDIX B

Subjective frequency estimation using judgements of 46 native speakers of BHA.

B-1: Frequency scale used by participants:



B-2: Frequency scale augmented for finer distinctions:

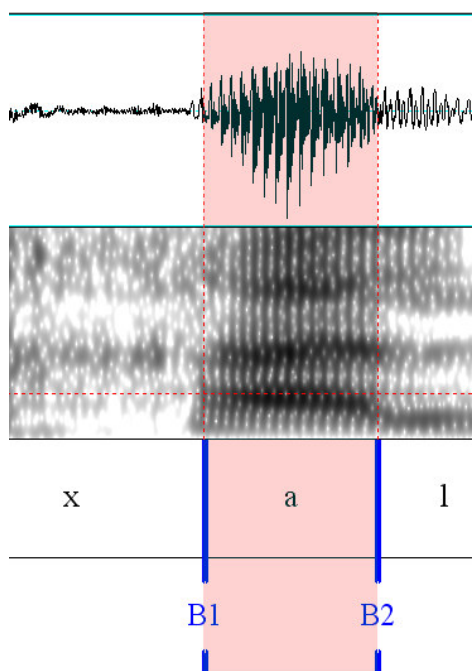


B-3: Conversion table (B-1 → B-2)

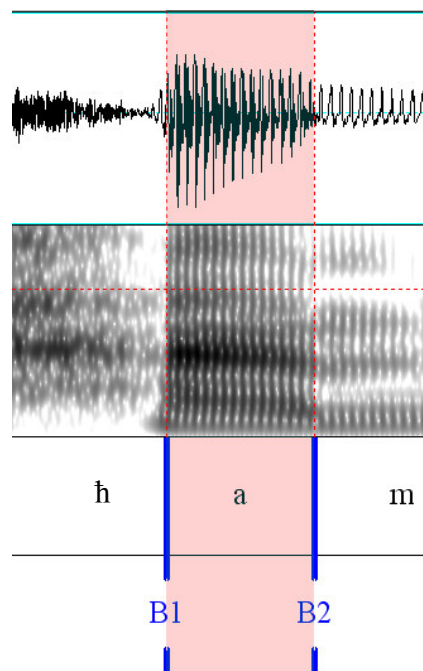
1 -1.33	1
1.34 - 1.66	2
1.67 - 2	3
2.01 - 2.33	4
2.34 -2.66	5
2.67-3	6

APPENDIX C

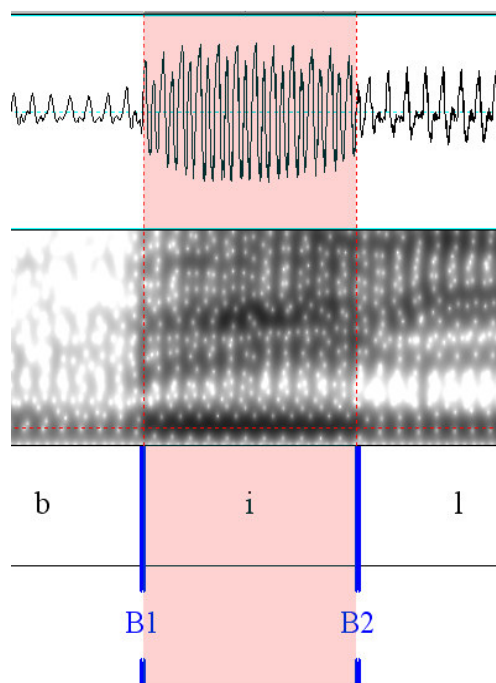
Examples of boundary locations in representative words produced by different speakers. Examples show waveforms and spectrograms with boundary marks (B1) and (B2) separating target vowels 'a' and 'i' from different consonants.



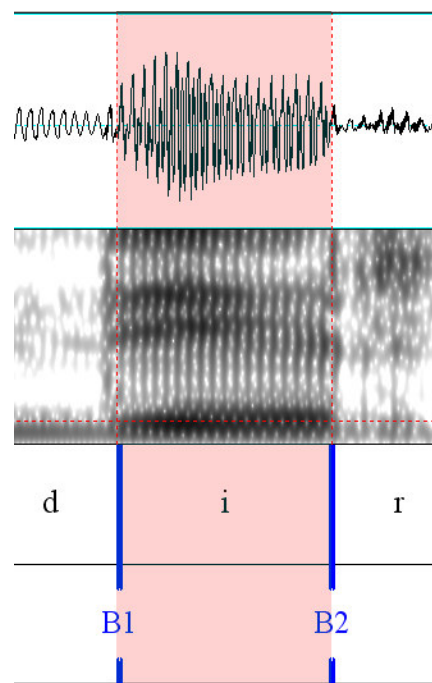
The word [daxal] by speaker L-A



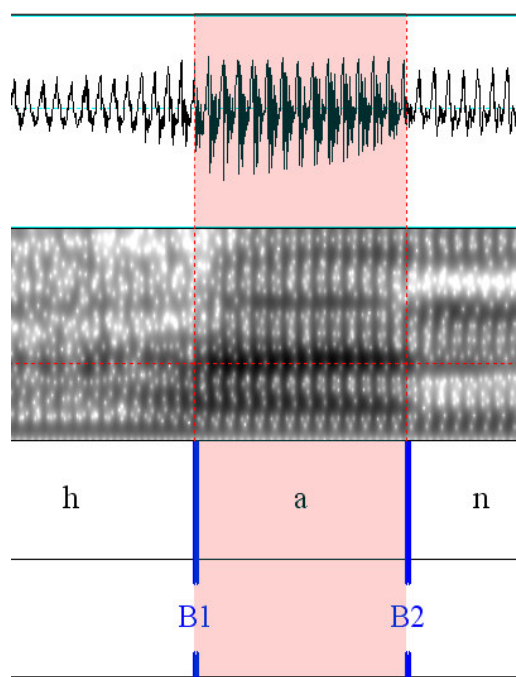
The word [laham] by Speaker L-B



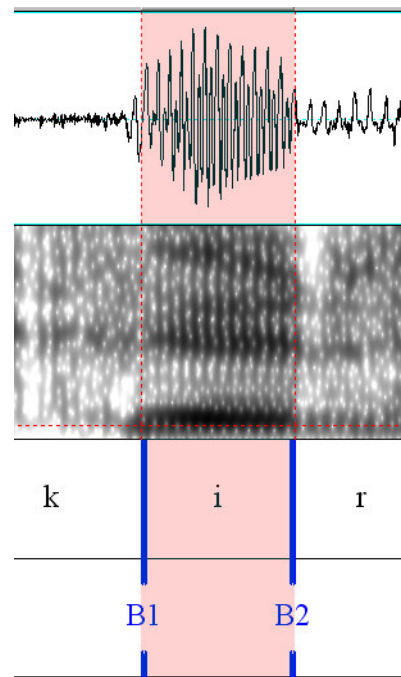
The word [gabil] by speaker L-C



The word [gidir] by Speaker L-D



The word [rahan] by speaker L-E

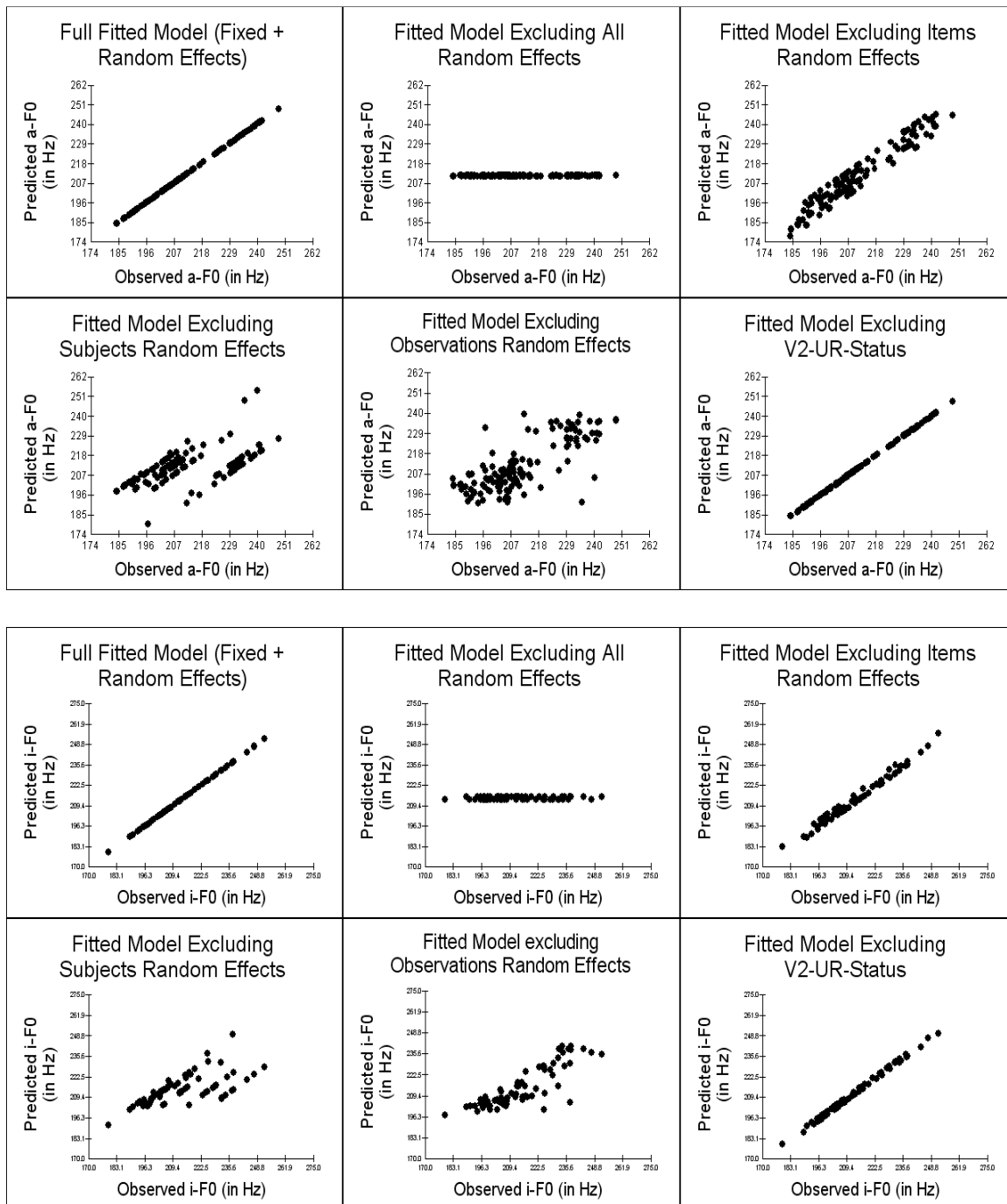


The word [fikir] by Speaker I-F

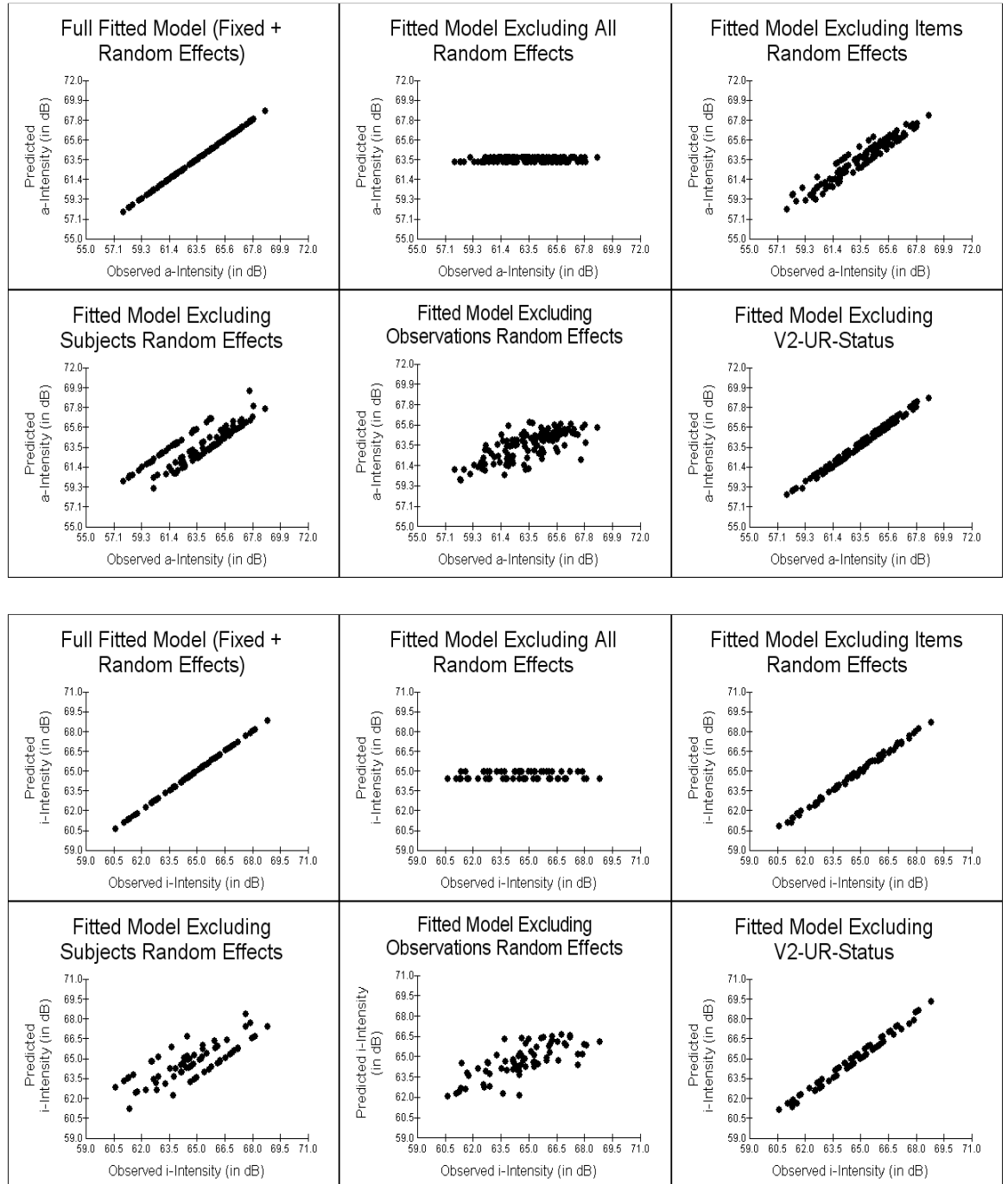
APPENDIX D

Prediction plots derived from multi-level models fitted to the BHA neutralisation data produced by seven native speakers. Fixed factors include an intercept and V2 Underlying Status (epenthetic vs lexical); random effects include items, speakers, and observations (i.e., renditions of each item by each speaker).

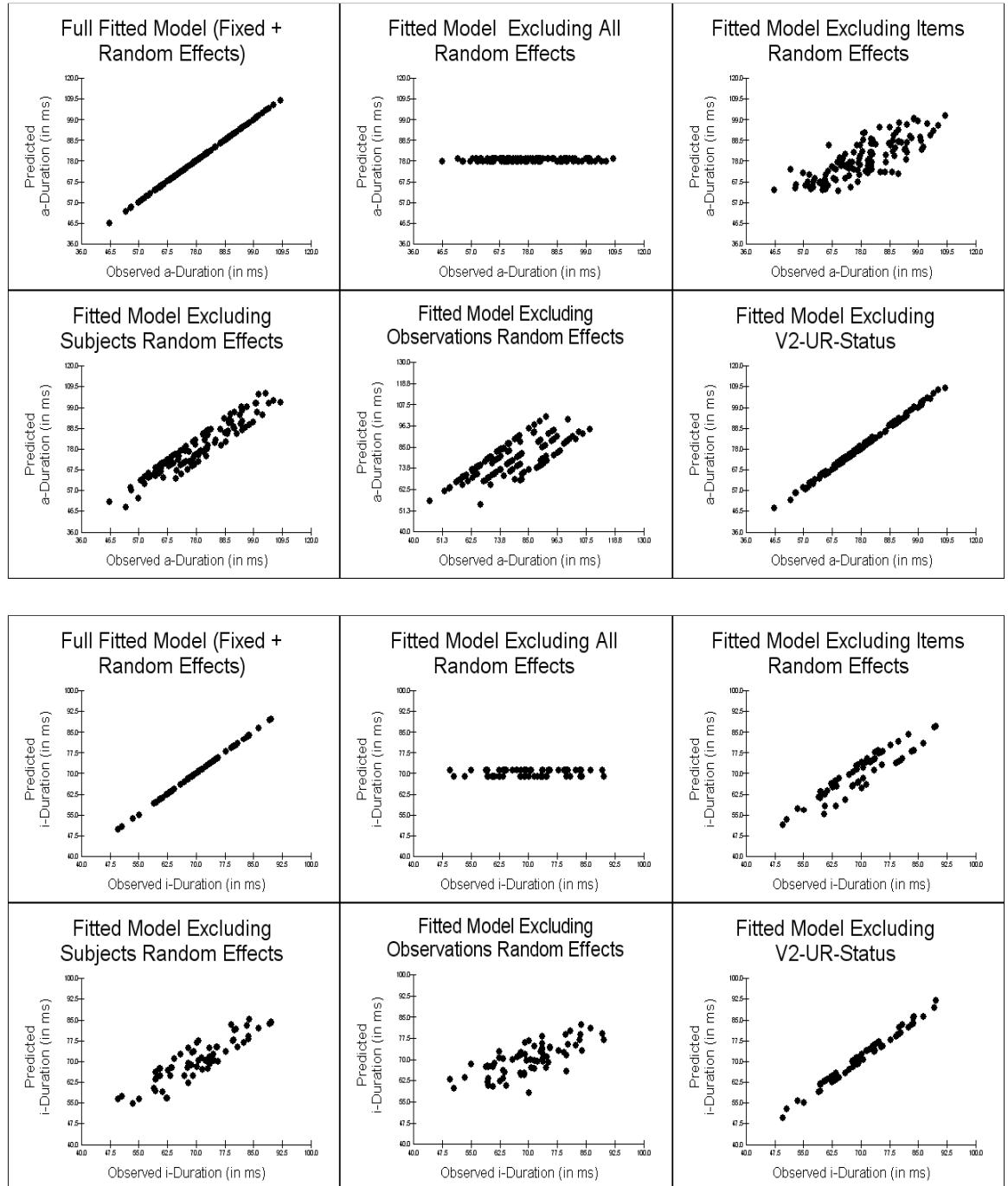
D-1: a-F0 and i-F0 data



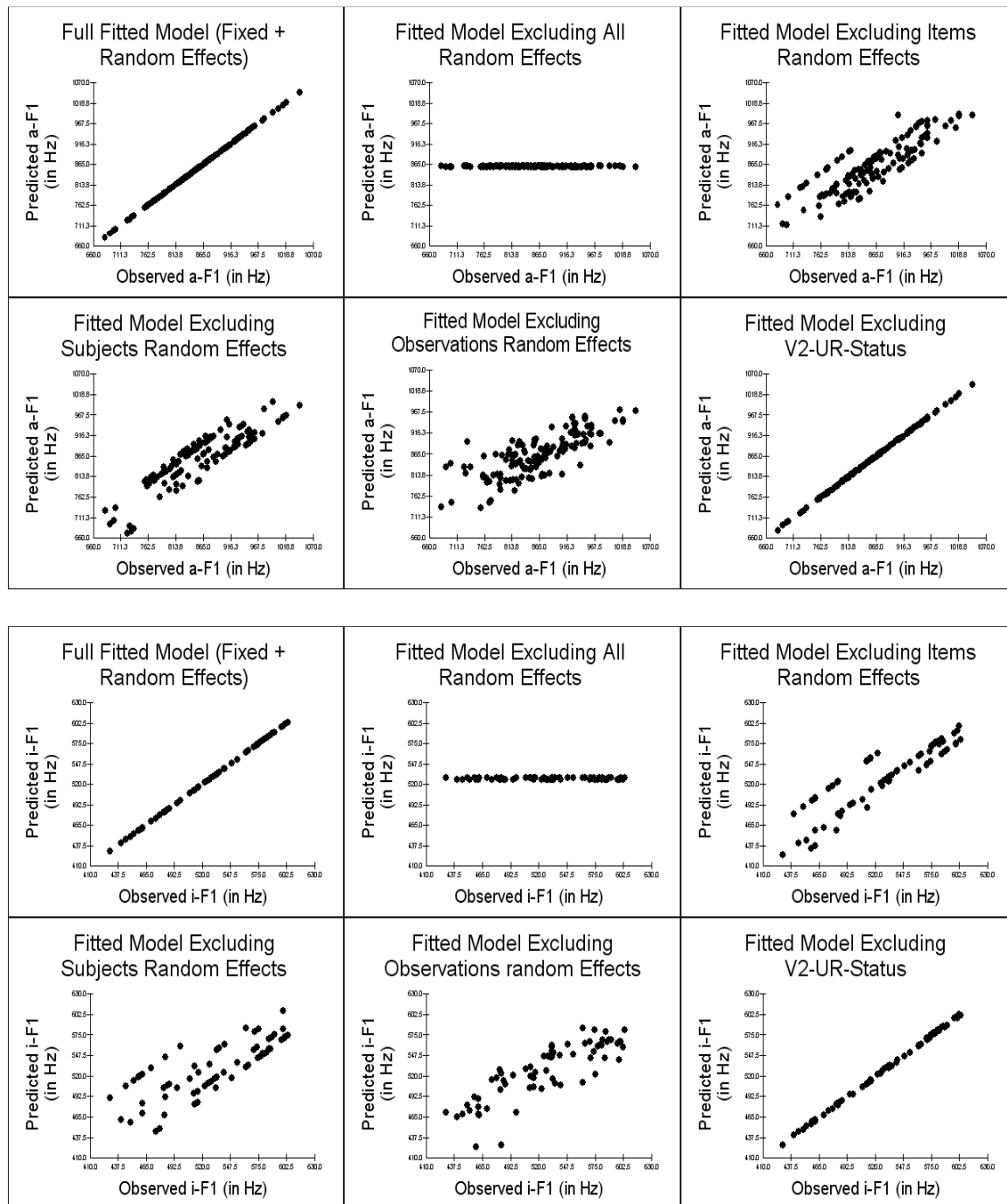
D-2: a-intensity and i-intensity data



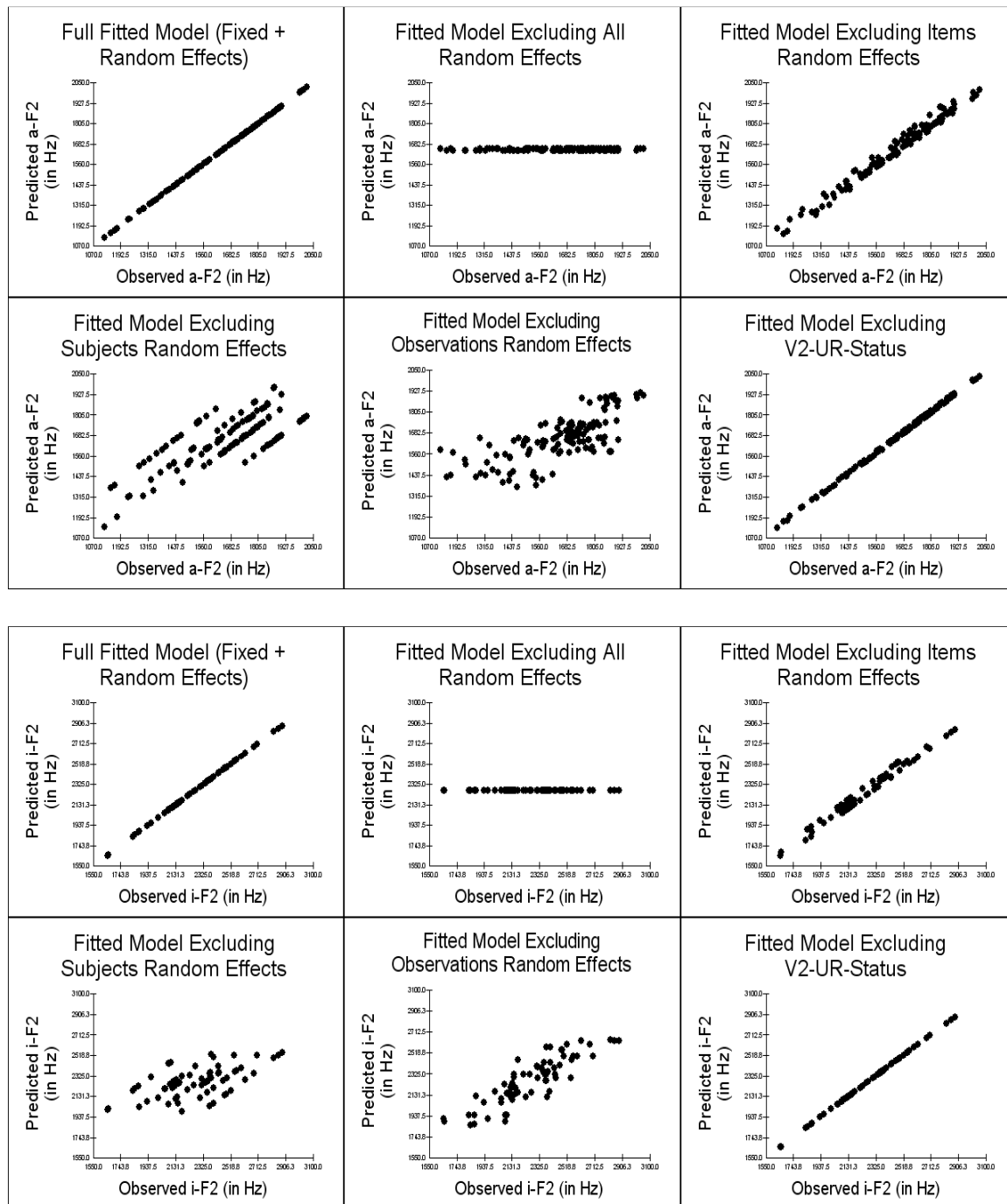
D-3: a-duration and i-duration data



D-4: a-F1 and i-F1 data



D-5: a-F2 and i-F2 data



APPENDIX E

Data used in the perception test (Speaker L-E):



APPENDIX F

Survey of 25 studies highlighting the correlation between the absence of minimal pairs in the stimuli and complete neutralisation. The inclusion criteria adopted are as follows. Studies mixing minimal pairs with non-minimal pairs or with nonsense minimal pairs in the test stimuli are excluded. Excluded too are studies exclusively using nonsense minimal pairs. With regard to orthography, only studies providing the orthographic form of the test words as part of the experiment procedures are included.

	<u>Orthography:</u>	<u>Stimuli:</u>	Number of studies reporting neutralisation to be	
	Contrast represented orthographically	Minimal pairs present	<u>(a): complete</u>	<u>(b): incomplete</u>
1	Yes	Yes	Kim & Jongman (1996)	Fougeron (2007) Fourakis & Port (1986) Fougeron & Steriade (1997) Fox & Terbeek (1977) Gerfen & Hall (2001) Jongman (2004) Dmitrieva et al (2010) Port & Crawford (1989) Port & O'Dell (1985) van Rooy et al (2003) Rudin (1980) Slowiczek & Dinnsen (1985) Snoeren et al (2006) Tieszen (1997) for Warsaw & for Bydgoszez Polish Warner et al (2004)
2	Yes	No	Arvaniti & Kilpatrick (2007) Lahiri & Hankamer (1988) Piroth & Janker (2004)	Cristófaró-Silva & Almeida (2008)
3	No	Yes	Warner et al (2006)	Charles-Luce & Dinnsen (1987)
4	No	No	Kopkallı (1993) Lahiri et al (1987)	

APPENDIX G

Mean and (SD) values of epenthetic and lexical vowels according to experimental conditions, stimuli, and tasks

G-1: a-F0 and i-F0 data (in Hz)

		TASK															
		Elicitation				Reading in context				Reading in a frame							
		[a]	/a/	[i]	/i/	[a]	/a/	[i]	/i/	[a]	/a/	[i]	/i/				
STIMULI	Members apart	Condition 1				Condition 2				Condition 3				Members apart			
	208.8 (15.5)	211 (15.6)	210 (11.7)	211.5 (9.8)	220.6 (24)	214.7 (17.8)	220 (26)	209.5 (15)	227.6 (18.8)	222.6 (12.4)	221 (13.6)	220 (14)	217 (19)	216 (15)	217 (17)	213.8 (13)	
STIMULI	Members close	Condition 4				Condition 5				Condition 6				Members close			
	210.7 (11.4)	210 (5.2)	207 (8)	207.6 (8)	211 (24.6)	212 (24.8)	211.6 (27)	212 (27)	215.9 (15.9)	215.9 (15.9)	215.9 (18.8)	215.5 (19)	212.7 (17)	212.7 (16)	211.6 (18.6)	211.7 (18.6)	
		Elicitation				Reading in context				Reading in a frame							
		209.7 (12.9)	210.7 (11)	208.6 (9.6)	209.6 (8.7)	215.9 (23.6)	213 (20)	215.8 (25.5)	210.8 (20.8)	218.7 (16.7)	219 (13.9)	218.5 (15.7)	217.9 (16)				

G-2: a-intensity and i-intensity data (in dB)

		TASK															
		Elicitation				Reading in context				Reading in a frame							
		[a]	/a/	[i]	/i/	[a]	/a/	[i]	/i/	[a]	/a/	[i]	/i/				
STIMULI	Members apart	Condition 1				Condition 2				Condition 3				Members apart			
	63.8 (1.6)	62.4 (2.2)	64.6 (1.4)	62.9 (1.7)	61.2 (1.2)	60.7 (1)	62.07 (1.2)	60.9 (1.5)	63.2 (3)	62.4 (2.4)	62.9 (1.2)	62.3 (.54)	62.7 (2.2)	61.8 (2)	63.2 (1.6)	62.07 (1.5)	
STIMULI	Members close	Condition 4				Condition 5				Condition 6				Members close			
	63.6 (2.3)	63.5 (2.4)	63.8 (2)	64.2 (1.8)	62.7 (4.8)	62.7 (5)	63.7 (4.2)	63.8 (4.1)	61.7 (2.8)	61.7 (2.7)	62.3 (2.4)	62.9 (2.7)	62.7 (3.3)	62.7 (3.4)	63.2 (2.9)	63.6 (2.8)	
		Elicitation				Reading in context				Reading in a frame							
		63.7 (1.9)	62.9 (2.2)	64.18 (1.6)	63.5 (1.8)	62 (3.4)	61.7 (3.6)	62.9 (3.1)	62.3 (3.3)	62.5 (2.8)	62.1 (2.4)	62.6 (1.8)	62.6 (1.8)				

G-3: a-duration and i-duration data (in ms)

		TASK															
		Elicitation				Reading in context				Reading in a frame							
		[a]	/a/	[i]	/i/	[a]	/a/	[i]	/i/	[a]	/a/	[i]	/i/	[a]	/a/	[i]	/i/
STIMULI	Members apart	Condition 1				Condition 2				Condition 3				Members apart			
		80.9 (8.4)	77.2 (5)	72.4 (5.4)	70 (5.7)	78.2 (9.4)	76.2 (7.5)	76.3 (7)	70.8 (6)	79.3 (6.8)	77.7 (7.6)	74 (5.3)	78 (6)	79.5 (7.7)	77 (6.3)	74 (5.8)	73 (6.7)
STIMULI	Members close	Condition 4				Condition 5				Condition 6				Members close			
		82.8 (8.4)	82.9 (8.2)	79.6 (10.5)	80 (6)	80.2 (13.6)	80.3 (12.7)	77.5 (12)	79 (11.9)	76.4 (6.4)	76.7 (6.7)	74 (5.7)	78.8 (5.5)	79.8 (9.6)	80 (9)	77 (9)	79 (7.7)
		Elicitation				Reading in context				Reading in a frame							
		81.9 (8)	80 (7)	75.9 (8.7)	75 (7.7)	79 (11)	78 (10)	76.9 (9)	74.9 (9.9)	77.9 (6.4)	77 (6.8)	74 (5)	78.5 (5.5)				

G-4: a-F1 and i-F2 data (in Hz)

		TASK															
		Elicitation				Reading in context				Reading in a frame							
		[a]	/a/	[i]	/i/	[a]	/a/	[i]	/i/	[a]	/a/	[i]	/i/	[a]	/a/	[i]	/i/
STIMULI	Members apart	Condition 1				Condition 2				Condition 3				Members apart			
		855 (52)	852 (42)	547 (37)	544 (29)	842 (62)	849 (64)	538 (39)	539 (34)	867 (71)	846 (81)	547 (39)	552 (32)	855 (58)	849 (60)	544 (36)	545 (30)
STIMULI	Members close	Condition 4				Condition 5				Condition 6				Members close			
		860 (71)	853 (75)	552 (36)	554 (34)	855 (100)	854 (90)	524 (43)	518 (43)	862 (92)	864 (100)	539 (40)	541 (48)	859 (82)	857 (82)	538 (39)	538 (42)
		Elicitation				Reading in context				Reading in a frame							
		857 (59)	853 (88)	549 (35)	549 (30)	849 (78)	851 (73)	531 (39)	529 (38)	864 (78)	855 (85)	543 (37)	546 (39)				

G-5: a-F2 and i-F2 data (in Hz)

		TASK															
		Elicitation				Reading in context				Reading in a frame							
		[a]	/a/	[i]	/i/	[a]	/a/	[i]	/i/	[a]	/a/	[i]	/i/	[a]	/a/	[i]	/i/
STIMULI	Members apart	Condition 1				Condition 2				Condition 3				Members apart			
		1707 (131)	1702 (117)	2391 (182)	2355 (169)	1616 (157)	1624 (144)	2415 (198)	2442 (161)	1554 (271)	1580 (202)	2432 (151)	2434 (197)	1625 (193)	1635 (156)	2413 (166)	2410 (168)
STIMULI	Members close	Condition 4				Condition 5				Condition 6				Members close			
		1663 (113)	1679 (91)	2417 (138)	2384 (126)	1558 (172)	1557 (159)	2462 (143)	2457 (175)	1585 (190)	1588 (173)	2398 (125)	2409 (95)	1602 (157)	1608 (145)	2426 (128)	2417 (130)
		Elicitation				Reading in context				Reading in a frame							
		1685 (118)	1691 (99)	2404 (153)	2369 (141)	1587 (158)	1590 (147)	2439 (164)	2450 (159)	1569 (221)	1584 (177)	2415 (132)	2422 (146)				

APPENDIX H

Mean paired differences and (SD) values of epenthetic and lexical vowels according to experimental conditions, stimuli, and tasks

H-1: a-F0 and i-F0 data (in Hz)

		TASK							
		Elicitation		Reading in context		Reading in a frame			
		[a]-/a/	[i]-/i/	[a]-/a/	[i]-/i/	[a]-/a/	[i]-/i/	[a]-/a/	[i]-/i/
STIMULI	Members apart	Condition 1		Condition 2		Condition 3		Members apart	
		-2.5 (8)	-1.5 (4.5)	6 (8.4)	10.4 (13)	-1 (11.9)	.6 (8.8)	.8 (9.7)	3.2 (10.3)
STIMULI	Members close	Condition 4		Condition 5		Condition 6		Members close	
		.57 (7)	-.35 (.76)	-.78 (1.5)	-.37 (3)	.03 (.8)	.5 (8)	-.06 (3.9)	-.08 (4.6)
		Elicitation		Reading in context		Reading in a frame			
		-.98 (7.3)	-.92 (3)	2.6 (6.7)	5 (10.6)	-.48 (8)	.54 (7.9)		

H-2: a-intensity and i-intensity data (in dB)

		TASK							
		Elicitation		Reading in context		Reading in a frame			
		[a]-/a/	[i]-/i/	[a]-/a/	[i]-/i/	[a]-/a/	[i]-/i/	[a]-/a/	[i]-/i/
STIMULI	Members apart	Condition 1		Condition 2		Condition 3		Members apart	
		1.4 (.7)	1.6 (1.4)	.56 (1.5)	1.2 (1.9)	.83 (1.7)	.64 (1.1)	.91 (1.3)	1.1 (1.5)
STIMULI	Members close	Condition 4		Condition 5		Condition 6		Members close	
		.09 (.63)	-.4 (.38)	-.04 (.62)	.04 (.35)	.03 (.33)	-.66 (.86)	.03 (.5)	-.37 (.6)
		Elicitation		Reading in context		Reading in a frame			
		.72 (.93)	.6 (1.5)	.26 (1.1)	.55 (1.4)	.43 (1.3)	-.01 (1.2)		

H-3: a-duration and i-duration data (in ms)

		TASK							
		Elicitation		Reading in context		Reading in a frame			
		[a]-/a/	[i]-/i/	[a]-/a/	[i]-/i/	[a]-/a/	[i]-/i/	[a]-/a/	[i]-/i/
STIMULI	members apart	Condition 1		Condition 2		Condition 3		Members apart	
		3.7 (5)	2.3 (5.5)	2 (2.6)	5.5 (3.5)	1.6 (5.5)	-4.2 (4.5)	2.4 (4.4)	1.2 (6)
STIMULI	members close	Condition 4		Condition 5		Condition 6		Members close	
		-1.1 (2)	-5.4 (5.7)	.05 (2)	-1.6 (4.3)	-.3 (3.9)	-4.8 (4.6)	-1.5 (2.6)	-2.3 (5)
		Elicitation		Reading in context		Reading in a frame			
		1.82 (4.2)	1 (5.5)	.95 (2.5)	1.9 (5.2)	.67 (4.6)	-4.5 (4.3)		

H-4: a-F1 and i-F1 data (in Hz)

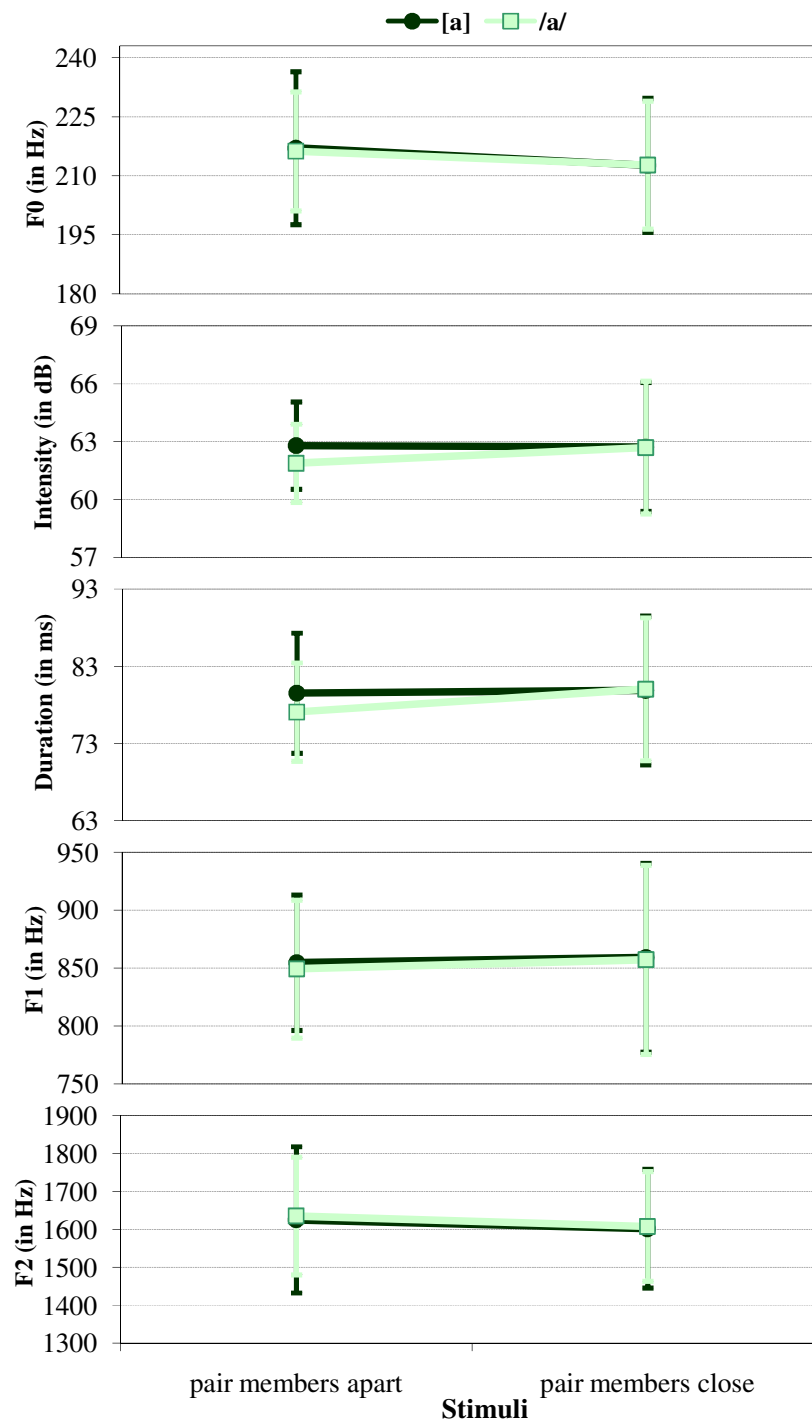
		TASK							
		Elicitation		Reading in context		Reading in a frame			
		[a]-/a/	[i]-/i/	[a]-/a/	[i]-/i/	[a]-/a/	[i]-/i/	[a]-/a/	[i]-/i/
STIMULI	members apart	Condition 1		Condition 2		Condition 3		Members apart	
		2.5 (15.5)	3 (22.5)	-6.6 (11)	-.84 (24.9)	20.5 (28)	-4.7 (13)	5.5 (21.5)	-.8 (19.5)
STIMULI	members close	Condition 4		Condition 5		Condition 6		Members close	
		7 (23.8)	-1.4 (9.6)	1 (11.7)	5.6 (3.4)	-2.3 (11.6)	-2 (18.6)	1.9 (16)	.7 (11.9)
		Elicitation		Reading in context		Reading in a frame			
		4.7 (19)	.85 (16.5)	-2.8 (11.4)	2.4 (17)	9.1 (23.5)	-3.4 (15)		

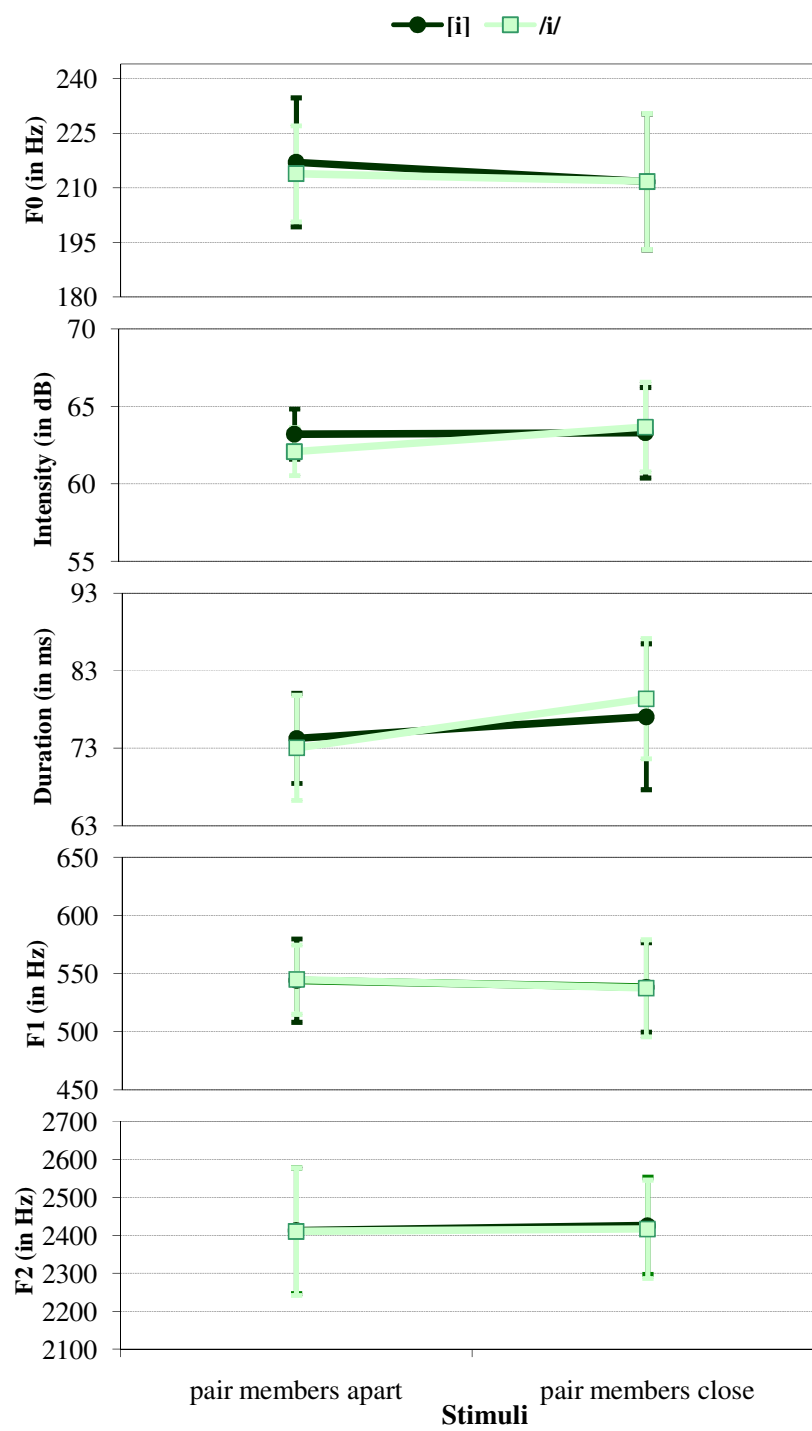
H-5: a-F2 and i-F2 data (in Hz)

		TASK							
		Elicitation		Reading in context		Reading in a frame			
		[a]-/a/	[i]-/i/	[a]-/a/	[i]-/i/	[a]-/a/	[i]-/i/	[a]-/a/	[i]-/i/
STIMULI	members apart	Condition 1		Condition 2		Condition 3		Members apart	
		4 (37)	36 (68.8)	-8 (76)	-27 (84.8)	-26.5 (118)	-2 (85)	-10 (79)	2.3 (78.8)
STIMULI	members close	Condition 4		Condition 5		Condition 6		Members close	
		-16.2 (61)	33 (45.7)	1.4 (37)	4.9 (62.6)	-3.3 (21.9)	-11.3 (87.5)	-6 (40.7)	8.9 (65.3)
		Elicitation		Reading in context		Reading in a frame			
		-6 (48.7)	34.6 (55)	-3.3 (56.7)	-11 (72)	-14.9 (81.3)	-6.6 (81.6)		

APPENDIX I

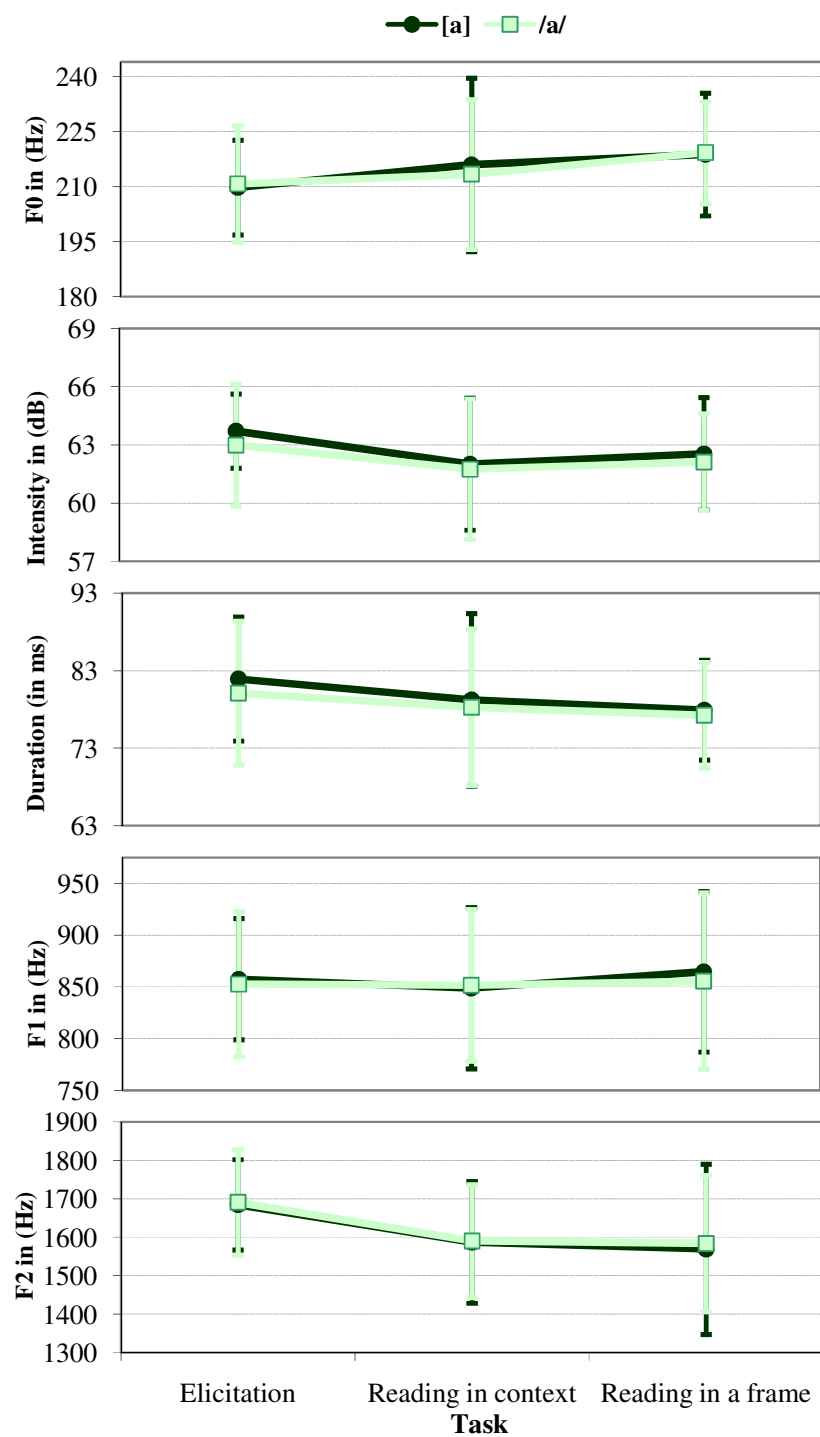
Main effects of stimulus list by V_2 Underlying Status; error bars show mean $\pm 1SD$

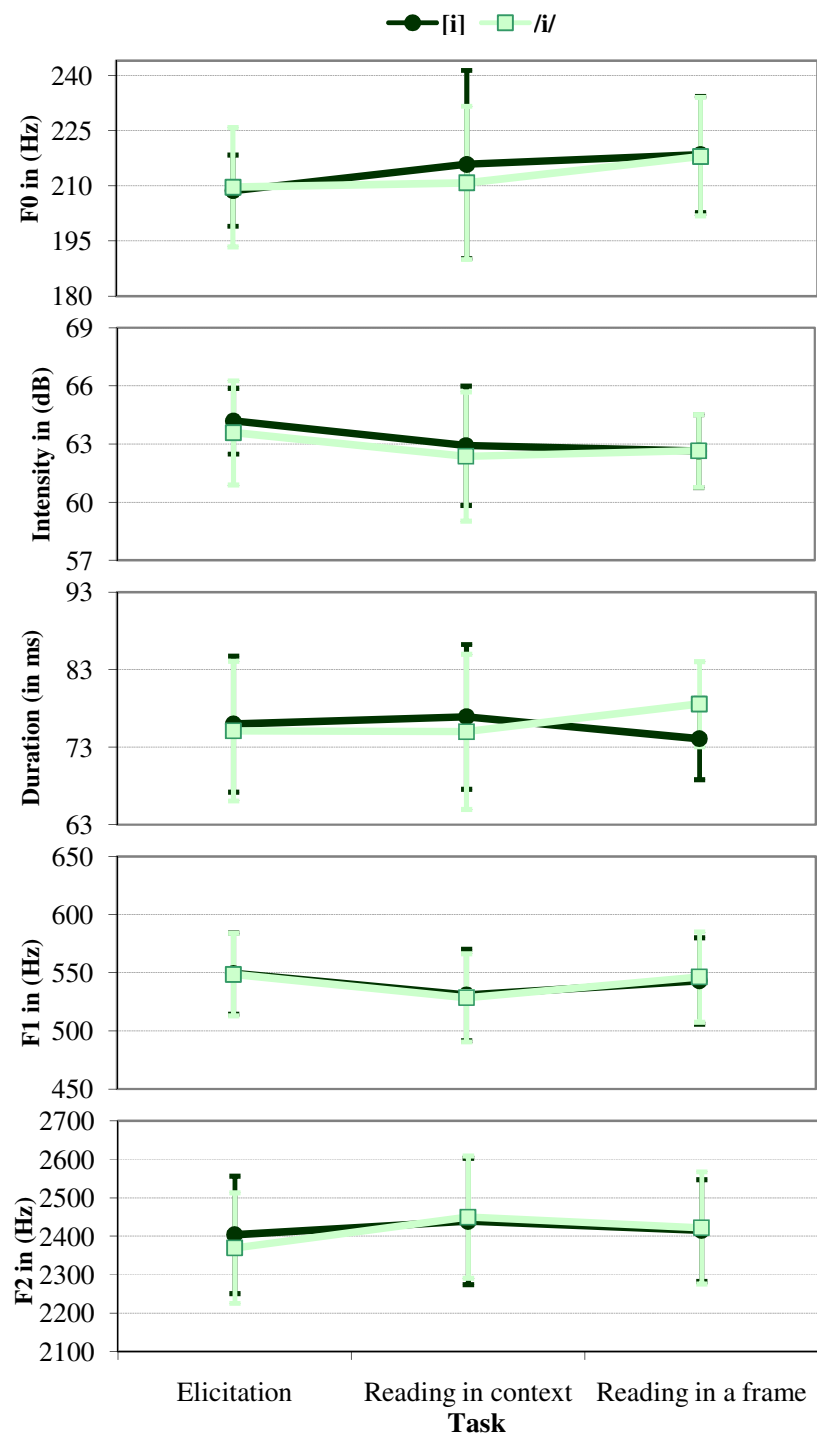




APPENDIX J

Main effects of task by V_2 Underlying Status. Error bars show mean $\pm 1SD$





REFERENCES

- Abelson, R. (1997). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). In L. Harlow, S. Mulaik & J. Steiger (eds.), *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum Associates. 117-141.
- Abercrombie, D. (1967). *Elements of general phonetics*. Edinburgh: Edinburgh University Press.
- Abu Bakar, H., K. Durvasula, N. Pincus & T. McKinnon (2007). Incomplete neutralization in KL Malay. Paper presented at *the 11th International Symposium on Malay/Indonesian Linguistics*. 6-8 August 2007, Manokwar, Papua.
- Abu-Mansour, M. (1992). Closed syllable shortening and morphological levels. In E. Broselow, M. Eid, & J. McCarthy (eds.), *Perspectives on Arabic linguistics IV: Papers from the fourth annual symposium on Arabic linguistics*. Amsterdam: John Benjamins. 47-75.
- Ahmed, A., E. Clarke & C. Adams (2008). Mismatch negativity and frequency representational width in children with specific language impairment. *Developmental Medicine and Child Neurology* 50: 938-944.
- Alderete, J. (1999). Head dependence in stress-epenthesis interaction: In B. Hermans & M. van Oostendorp (eds.), *The derivational residue in phonological optimality theory*. Amsterdam: John Benjamins. 29-50.
- Alderete, J. & B. Tesar (2002). Learning covert phonological interaction: An analysis of the problem posed by the interaction of stress and epenthesis. *Report no. RuCCS-TR-72*, Piscataway, NJ: Rutgers Centre for Cognitive Science.
- Alfonso, P. & T. Baer (1982). Dynamics of vowel articulation. *Language and Speech* 25: 151-173.
- Alghamdi, M. (1998). A spectrographic analysis of Arabic vowels: A cross-dialect study. *Journal of King Saud University: Arts* 10, 1: 3-24.
- Alghamdi, M. (2004). *Analysis, synthesis and perception of voicing in Arabic*. Riyadh: Al-Toubah.
- Al-Hazmy, A. (1972). *A critical and comparative study of the spoken dialects of Badr and District in Saudi Arabia*. MPhil thesis, University of Leeds.
- Al-Hazmy, A. (1975). *A critical and comparative study of the spoken dialect of the Harb tribe in Saudi Arabia*. PhD thesis, University of Leeds.
- Allen, J. & J. Miller (2001). Contextual influences on the internal structure of phonetic categories: A distinction between lexical status and speaking rate. *Perception and Psychophysics* 63, 5: 798-810.
- Allen, J. & J. Miller (2004). Listener sensitivity to individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America* 115, 6: 3171-3183.
- Almihmadi, M. (2006). *Vowel intrusion: Evidence from a Bedouin Hijazi Arabic Dialect*. MA dissertation, University College London.

- Al-Mozainy, H. (1981). *Vowel alternations in a Bedouin Hijazi Arabic dialect: Abstractness and stress*. PhD thesis, University of Texas at Austin.
- Al-Mozainy, H., R. Bley-Vorman & J. McCarthy (1985). Stress shift and metrical structure. *Linguistic Inquiry* 16: 135-144.
- Al-Tamimi, J. & E. Ferragne (2005). Does vowel space size depend on language vowel inventories? Evidence from two Arabic dialects and French. *Proceedings of Interspeech*. 4-8 September 2005, Lisbon. 2465-2468.
- Altman, D., D. Machin, T. Bryant & M. Gardner (2000). *Statistics with confidence: Confidence intervals and statistical guidelines*. 2nd edition. Bristol: BMJ.
- Amenedo, E. & C. Escera (2000). The accuracy of sound duration representation in the human brain determines the accuracy of behavioural perception. *European Journal of Neuroscience* 12: 2570-2574.
- Archangeli, D. (1984). *Underspecification in Yawelmani phonology and morphology*. PhD thesis, MIT.
- Archangeli, D. (1988). Aspects of underspecification theory. *Phonology* 5: 183-207.
- Archangeli, D. & D. Pulleyblank (1994). *Grounded phonology*. Cambridge, MA: MIT Press.
- Armstrong, J. (2007a). Significance tests harm progress in forecasting. *International Journal of Forecasting* 23: 321-327.
- Armstrong, J. (2007b). Statistical significance tests are unnecessary even when properly done and properly interpreted: Reply to commentaries. *International Journal of Forecasting* 23, 2: 335-336.
- Arvaniti, A. (2006). Stop epenthesis revisited. Poster presented at *the 10th Conference on Laboratory Phonology (LabPhon 10)*. 30 June-2 July 2006, Paris: Université de Paris.
- Arvaniti, A. & C. Kilpatrick (2007). The production and perception of epenthetic stops. Talk presented at the *LSA meeting*. 4-7 January 2007, Anaheim, CA.
- Baayen, R. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.
- Baayen, R., D. Davidson & D. Bates (2008). Mixed-effects modelling with crossed random effects for subjects and items. *Journal of Memory and Language* 59, 4: 390-412.
- Baese, M. & M. Goldrick (2006). Lexical effects on phonetic variation independent of phonotactics. Poster presented at *the 10th Conference on Laboratory Phonology (LabPhon 10)*. 30 June-2 July 2006, Paris: Université de Paris.
- Baese, M., T. Poepsel & M. Goldrick (2007). Moving new words into the neighborhood: A preliminary report. Poster presented at *the Fourth International Workshop on Language Production*. 3-5 September 2007, Münster, Germany.
- Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin* 66, 6: 423-437.
- Baković, E. (2007). A revised typology of opaque generalizations. *Phonology* 24: 217-259.
- Balota, D., M. Pilotti & M. Cortese (2001). Subjective frequency estimates for 2,938 monosyllabic words. *Memory and Cognition* 29: 639-647.

- Barcroft, J. & M. Sommers (2005). Effects of acoustic variability on second language vocabulary learning. *Studies in Second Language Acquisition* 24: 387-414.
- Barnes, J. (2006). *Strength and weakness at the interface: Positional neutralization in phonetics and phonology*. Berlin: Mouton de Gruyter.
- Barnes, J. (submitted). *Phonetics and Phonology in Russian unstressed vowel reduction: A study in hyperarticulation*. <http://www.bu.edu/linguistics/UG/barnes/>.
- Baroni, M. & L. Vanelli (2000). The relationship between vowel length and consonantal voicing in Friulian. In L. Repetti (ed.), *Phonological theory and the dialects of Italy*. Amsterdam: John Benjamins. 13-44.
- Basalyga, G. & E. Salinas (2006). When response variability increases neural network robustness to synaptic noise. *Neural Computation* 18: 1349-1379.
- Baumann, M. (1995). *The production of syllables in connected speech*. PhD thesis, University of Nijmegen.
- Bentin, S. & R. Ibrahim (1996). New evidence for phonological processing during visual word recognition: The case of Arabic. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22, 2: 309-323.
- Benzeghiba, M., R. De Mori, O. Derou, S. Dupont, T. Erbes, D. Jouvét, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi & C. Wellekens (2007). Automatic speech recognition and speech variability: A review. *Speech Communication* 49: 763-786.
- van Bergem, D. (1995). *Acoustic and lexical vowel reduction*. Amsterdam: Dordrecht ICG Printing.
- Berger, J. & T. Selke (1987). Testing a point null hypothesis: The irreconcilability of p-values and evidence. *Journal of American Statistical Association* 82: 112-139.
- Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association* 33, 203: 526-536.
- Bermúdez-Otero, R. (1999). *Constraints interaction in language change: Quantity in English and German*. PhD thesis, University of Manchester.
- Bickel, D. (2002). Robust estimators of the mode and skewness of continuous data. *Computational Statistics and Data Analysis* 39: 153-163.
- Bickel, D. (2003). Robust and efficient estimation of the mode of continuous data: The mode as a viable measure of central tendency. *Journal of Statistical Computation and Simulation* 73: 899-912.
- Bickel, D. & R. Frühwirth (2006). On a fast, robust estimator of the mode: Comparisons to other robust estimators with applications. *Computational Statistics and Data Analysis* 50: 3500-3530.
- Bird, K., D. Hadzi-Pavlovic & A. Isaac (2000). *PSY: A program for contrast analysis*. <http://www.psy.unsw.edu.au/research/PSY.htm>.
- Blache, P. & C. Meunier (2004). Language as a complex system: The case of phonetic variability. *VI Congreso de Lingüística General*. 3-7 May 2004, Santiago de Compostela.
- Bladon, R. & A. Al-Bamerni (1976). Coarticulation resistance of English /l/. *Journal of Phonetics* 4: 137-150.
- Bladon, R. & F. Nolan (1977). A video-fluorographic investigation of tip and blade alveolars in English. *Journal of Phonetics* 5: 185-193.

- Blanc, H. (1970). The Arabic dialect of the Negev Bedouins. *Proceedings of the Israel academy of sciences and humanities* 4: 112–50.
- Blevins, J. & A. Garrett (1998). The origins of consonant-vowel metathesis. *Language* 74, 3: 508-556.
- Blumstein, S. (1986). On acoustic invariance in speech. In J. Perkell & D. Klatt (eds.), *Invariance and Variability in Speech Processes*. Hillsdale NJ: Lawrence Erlbaum. 178-201.
- Blumstein, S. (1991). The relation between phonetics and phonology. *Phonetica* 48: 108–119.
- Boersma, P. & D. Weenink (2008). *Praat: doing phonetics by Computer*. [Computer Programme: available at <http://www.praat.org>].
- Boly, M., E. Balteau, C. Schnakers, C. Degueldre, G. Moonen, A. Luxen, C. Phillips, P. Peigneux, P. Maquet & S. Laureys (2007). Baseline brain activity fluctuations predict somatosensory perception in humans. *Proceedings of the National Academy of Sciences* 104, 29: 12187-12192.
- Bonneau, A. (1996). Identification of vowel features from French stop bursts. In *the Fourth International Conference on Spoken Language Processing-ICSLP'96*. Philadelphia. 2506 - 2509.
- Boomershine, A. (2005). *Perceptual processing of variable input in Spanish: An exemplar-based approach to speech perception*. PhD thesis, The Ohio State University.
- Bouchhioua, N. (2008). Duration as a cue to stress and accent in Tunisian Arabic, native English, and L2 English. In *SP-2008*. 535-538.
- Bouchhioua, N. (2009). Stress and accent in Tunisian Arabic. Paper presented at *the First International Conference on Intonational Variation in Arabic*. 28-29 September 2009, York: University of York.
- Boyce, S. (1990). Coarticulatory organization for lip rounding in Turkish and English. *Journal of the Acoustical Society of America* 88: 2584-2595.
- Bradlow, A., L. Nygaard & D. Pisoni (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception and Psychophysics* 61, 2: 206-219.
- Bradlow, A. & D. Pisoni (1999). Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors. *Journal of the Acoustical Society of America* 106: 2074–2085.
- Brannon, E., S. Abbott & D. Lutz (2004). Number bias for discrimination of large visual sets in infancy. *Cognition* 93: B59-B68.
- Braver, A. (2010a). Is there incomplete neutralization in American English flapping. Poster presented at *the 18th Manchester Phonology Meeting (MFM)*. 20-22 May 2010, Manchester: University of Manchester.
- Braver, A. (2010b). Incomplete neutralisation in American English flapping: A production study. *Pennsylvania Linguistics Colloquium (PLC)* 34. Pennsylvania University.
- Brockhaus, W. (1995). *Final devoicing in the phonology of German*. Tübingen: Max Niemeyer Verlag.

- Broselow, E. (1982). On predicting the interaction of stress and epenthesis. *Glossa: An International Journal of Linguistics* 16, 2: 115-132.
- Browman, C. & L. Goldstein (1989). Articulatory gestures as phonological units. *Phonology* 6: 201-251.
- Buchwald, A. (2005). *Sound structure preservation, repair and well-formedness: grammar in spoken language production*. PhD thesis, Johns Hopkins University.
- Burton, M. & K. Robblee (1997). A phonetic analysis of voicing assimilation in Russian. *Journal of Phonetics* 25: 97-114.
- Butcher, A. (1995). The phonetics of neutralisation: The case of Australian coronals. In J. Windsor Lewis (ed.), *Studies in general and English phonetics. Essays in honour of J.D. O'Connor*. London: Routledge. 10-38.
- Butcher, A. & K. Ahmad (1987). Some acoustic and aerodynamic characteristics of pharyngeal consonants in Iraqi Arabic. *Phonetica* 44: 156-172.
- Calabrese, A. & W. Wetzels (eds.) (2009). *Loan phonology*. Amsterdam: John Benjamins.
- Campbell, L. (1974). Theoretical implications of Kekchi phonology. *International Journal of American Linguistics* 40, 4: 269-278.
- Campos-Astorkiza, R. (2007). *Minimal contrast and the phonology-phonetics interaction*. PhD thesis, University of Southern California.
- Campos-Astorkiza, R. (2008). Two sources of voicing neutralization in Lithuanian. In A. Bonitis (ed.), *Proceedings of the International Speech Communication Association Research Workshop on Experimental Linguistics*. Athens, Greece: University of Athens. 49-53.
- Carlyon, R. & B. Moore (1984). Intensity discrimination: A severe departure from Weber's law. *Journal of the Acoustical Society of America* 76: 1369-1376.
- Carroll, J. (1971). Measurement properties of subjective magnitude estimates of word frequency. *Journal of Verbal Learning and Verbal Behavior* 10: 722-729.
- Carroll, S. (2006). Saliency, awareness and SLA. In M. O'Brien, C. Shea & J. Archibald (eds.), *Proceedings of the 8th Generative Approaches to Second Language Acquisition Conference (GASLA 2006)*. Somerville, MA: Cascadia Proceedings Project. 17-24.
- Carver, R. (1978). The case against statistical significance testing. *Harvard Educational Review* 48: 378-399.
- Charles-Luce, J. (1985). Word final devoicing in German: Effects of phonetic and sentential contexts. *Journal of Phonetics* 13: 309-324.
- Charles-Luce, J. (1986). Assimilatory versus nonassimilatory neutralization processes in Catalan. *Journal of the Acoustical Society of America* 79, Supplement 1: S26.
- Charles-Luce, J. (1993). The effects of semantic context on voicing neutralization. *Phonetica* 50: 28-43.
- Charles-Luce, J. & D. Dinnsen (1987). A reanalysis of Catalan devoicing. *Journal of Phonetics* 15: 187-190.
- Chatfield, C. (1995). *Problem solving: A statistician's guide*. 2nd edition. Boca Raton: Chapman and Hall/CRC.
- Chen, M. (1970). Vowel length variation as a function of the voicing of the consonant environment. *Phonetica* 22: 129-159.
- Cho, T. (2002). *The effects of prosody on articulation in English*. New York: Routledge.

- Cho, T. (2004). Prosodically conditioned strengthening and vowel-to-vowel coarticulation in English. *Journal of Phonetics* 32: 141–176.
- Cho, T. & J. McQueen (2006). Phonological versus phonetic cues in native and non-native listening: Korean and Dutch listeners' perception of Dutch and English consonants. *Journal of the Acoustical Society of America* 119, 5: 3085-3096.
- Clark, H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior* 12: 335-359.
- Clements, G. (1990). The role of the sonority cycle in core syllabification. In J. Kingston & M. Beckman (eds.), *Papers in laboratory phonology I: Between the grammar and physics of speech*. Cambridge: Cambridge University Press. 283-333.
- Cleveland, W. (1985). *The elements of graphing data*. New York: Chapman and Hall.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. 2nd edition. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist* 49, 12: 997-1003.
- Cole, R., M. Coltheart & F. Allard (1974). Memory of a speaker's voice: Reaction time to same- or different-voiced letters. *The Quarterly Journal of Experimental Psychology* 26: 1-7.
- Coleman, J. (2003). Discovering the acoustic correlates of phonological contrasts. *Journal of Phonetics* 31: 351–372.
- Colley, M. (2009). Investigating the incomplete neutralization of flaps in North American English. *Journal of the Acoustical Society of America* 126, 4: 2181.
- Connine, C. & E. Pinnow (2006). Phonological variation in spoken word recognition: Episodes and abstractions. *The Linguistic Review* 23: 235–245.
- Conwell, E. & J. Morgan (2007). Resolving grammatical category ambiguity in Acquisition. In H. Caunt-Nulton, S. Kulatilake & I-h. Woo (eds.), *Proceedings of the 31st Annual Boston University Conference on Language Development*. 117-128.
- Creel, S., R. Aslin & M. Tanenhaus (2008). Heeding the voice of experience: The role of talker variation in lexical access. *Cognition* 106: 633-664.
- Cristófaró-Silva, T. & L. Almeida (2008). On the nature of epenthetic vowels. In L. Bisol & C. Brescancini (eds.), *Contemporary phonology in Brazil*. Cambridge: Cambridge Scholars Publishing. 193-212.
- Cullinan, W. & M. Tekieli (1979). Perception of vowel features in temporally-segmented noise portions of stop consonant CV syllables. *Journal of Speech and Hearing Research* 22: 122–131.
- Daniel, L. (1998). Statistical significance testing: A historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. *Research in the Schools* 5, 2: 23-32.
- Davidson, L. (2007). The relationship between the perception of non-native phonotactics and loanword adaptation. *Phonology* 24: 261-286.
- Davidson, L. & K. Roon (2008). Durational correlates for differentiating consonant sequences in Russian. *Journal of the International Phonetic Association* 38: 137-165.
- de Jong, K. & B. Zawaydeh (1999). Stress, duration, and intonation in Arabic word-level prosody. *Journal of Phonetics* 27: 3-22.

- de Lacy, P. (2002a). *The formal expression of markedness*. Ph.D. dissertation, University of Massachusetts, Amherst.
- de Lacy, P. (2002b). The interaction of tone and stress in Optimality Theory. *Phonology* 19: 1-32.
- delMas, R. & Y. Liu (2004). Students' understanding of factors that affect the standard deviation. Paper presented at *ICME-10 Denmark, TSG 11: Research and development in the teaching and learning of probability and statistics*. 4-11 July 2004, Copenhagen. Retrieved 30 May 2009 from <http://www.icme-organisers.dk/tsg11/>.
- Diehl, R. & K. Kluender (1987). On the categorization of speech sounds. In S. Harnad (ed.), *Categorical perception: The groundwork of cognition*. Cambridge: Cambridge University Press. 226-253.
- Dinnsen, D. (1985). A re-examination of phonological neutralization. *Journal of Linguistics* 21: 265-279.
- Dinnsen, D. & J. Charles-Luce (1984). Phonological neutralization, phonetic implementation, and individual differences. *Journal of Phonetics* 12: 49-60.
- Dmitrieva, O., A. Jongman & J. Sereno (2010). Phonological neutralization by native and non-native speakers: The case of Russian final devoicing. *Journal of Phonetics* 38, 3: 483-492.
- Donaldson, W. (1992). Measuring recognition memory. *Journal of Experimental Psychology: General* 121, 3: 275-277.
- Duch, W. (1996). Categorization, prototype theory and neural dynamics. In T. Yamakawa & G. Matsumoto (eds.), *Proceedings of SoftComputing'96*, Iizuka, Japan. 482-485.
- Dufour, S., N. Nguyen & U. Frauenfelder (2007). The perception of phonemic contrasts in a non-native dialect. *Journal of the Acoustical Society of America* 121: EL131-EL136.
- Dufour, S., N. Nguyen & U. Frauenfelder (2010). Does training on a phonemic contrast absent in the listener's dialect influence word recognition? *Journal of the Acoustical Society of America* 127, 6: EL1-EL6.
- Dunlap, W., J. Cortina, J. Vaslow & M. Burke (1996). Meta-analysis of experiments with matched groups of repeated measures designs. *Psychological Methods* 1, 2: 170-177.
- Eddy, D. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic & A. Tversky (eds.), *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press. 249-267.
- Eimas, P. & J. Corbit (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology* 4: 99-109.
- Elman, J. (1992). Connectionist approaches to acoustic/phonetic processing. In W. Marslen-Wilson (ed.), *Lexical representation and process*. 2nd edition. Cambridge, MA: MIT Press. 227-260.
- Elman J. & J. McClelland (1986). Exploiting lawful variability in the speech wave. In J. Perkell & D. Klatt (eds.), *Invariance and variability in speech processes*. London: LEA. 360-381.
- Ernestus, M. (in press). Gradience and categoricity in phonological theory. In M. van Oostendorp, C. Ewen, B. Hume, and K. Rice (eds.), *The Blackwell companion to phonology*. Oxford: Wiley-Blackwell.

- Ernestus, M. & H. Baayen (2006). The functionality of incomplete neutralization in Dutch: the case of past-tense formation. In L. Goldstein, D. Whalen & C. Best (eds.), *Laboratory phonology 8*. Berlin: Mouton de Gruyter. 27–49.
- Ettlinger, M. (2007). Incomplete neutralisation in Dutch opacity. *Journal of the Acoustical Society of America* 122, 5: 2997.
- Eulitz, C. & A. Lahiri (2004). Neurobiological evidence for abstract phonological representations in the mental lexicon during speech recognition. *Journal of Cognitive Neuroscience* 16, 4: 577-583.
- Eviatar, Z., R. Ibrahim & D. Ganayim (2004). Orthography and the hemispheres: Visual and linguistic aspects of letter processing. *Neuropsychology* 18, 1: 174-184.
- Farwaneh, S. (1995). *Directionality effects in Arabic dialect syllable structure*. PhD thesis, University of Utah.
- Finley, S. (2008). *Formal and cognitive restrictions on vowel harmony*. PhD thesis, Johns Hopkins University.
- Finley, S. (2009). *The interaction of vowel harmony and epenthesis*. Manuscript, Johns Hopkins University.
- Fischer-Jørgensen, E. (1990). Intrinsic F0 in tense and lax vowels with special reference to German. *Phonetica* 47: 99-140.
- Fisher, W. & I. Hirsh (1976). Intervocalic flapping in English. In S. Mufwene, C. Walker & S. Steever (eds.), *Papers from the 12th Regional Meeting, Chicago Linguistic Society*. Chicago: Chicago Linguistic Society. 183-198.
- Flanagan, J. (1955a). Difference limen for the intensity of a vowel sound. *Journal of the Acoustical Society of America* 27, 6: 1223-1225.
- Flanagan, J. (1955b). A difference limen for vowel formant frequency. *Journal of the Acoustical Society of America* 27, 3: 613-617.
- Flanagan, J. (1957). Estimates of the maximum precision necessary in quantizing certain 'dimensions' of vowel sounds. *Journal of the Acoustical Society of America* 29: 533-534.
- Flanagan, J. & M. Saslow (1958). Pitch discrimination for synthetic vowels. *Journal of the Acoustical Society of America* 30, 5: 435-442.
- Flemming, E. (1997). Phonetic Optimization: Compromise in Speech Production. *University of Maryland Working Papers in Linguistics 5: Selected phonology papers from H-OT-97*.
- Florentine, M., S. Buus, B. Scharf & G. Canevet (1984). Speech reception thresholds in noise for native and non-native listeners. *Journal of the Acoustical Society of America* 75, S1: S84.
- Fougeron, C. (2001). Articulatory properties of initial segments in several prosodic constituents in French. *Journal of Phonetics* 29: 109-135.
- Fougeron, C. (2007). Word boundaries and contrast neutralization in the case of enchaînement in French. In J. Cole & J. Hualde (eds.), *Papers in laboratory phonology IX: Change in Phonology*. Berlin: Mouton de Gruyter. 609-642.
- Fougeron, C. & D. Steriade (1997). Does deletion of French schwa lead to neutralization of lexical distinctions? *Actes Eurospeech'97*. Rhodes, Greece.
- Fourakis, M. (1980). A phonetic study of sonorant-fricative clusters in two dialects of English. *Research in Phonetics* 1: 167-200.

- Fourakis, M. (1984). Should neutralization be redefined? *Journal of Phonetics* 12: 291-296.
- Fourakis, M. & G. Iverson (1984). On the 'incomplete neutralization' of German final obstruents. *Phonetica* 41: 140-149.
- Fourakis, M. & R. Port (1986). Stop epenthesis in English. *Journal of Phonetics* 14: 197-221.
- Fowler, C. & J. Housum (1987). Talkers' signaling of 'new' and 'old' words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language* 26: 489-504.
- Fox, R. & D. Terbeek (1977). Dental flaps, vowel duration, and rule ordering in English. *Journal of Phonetics* 5: 27-34.
- Freedman, D. & P. Diaconis (1981). On the histogram as a density estimator: L2 theory. *Probability Theory and Related Fields* 57, 4: 453-476.
- Fry, D. (1958). Experiments in the perception of stress. *Language and Speech* 1: 126-152.
- Fujisaki, H., K. Nakamura & T. Imoto (1975). Auditory perception of duration of speech and nonspeech stimuli. In G. Fant & M. Tatham (eds.), *Auditory analysis and perception of speech*. New York: Academic Press. 197-220.
- Gafos, A. (2003a). *Dynamics: The non-derivational alternative to modelling phonetics-phonology*. Manuscript, New York University.
- Gafos, A. (2003b). Greenberg's asymmetry in Arabic: A consequence of stems in paradigms. *Language* 79: 317-355.
- Gafos, A. (2006). Dynamics in grammar: Comments on Ladd and Ernestus & Baayen. In L. Goldstein, D. Whalen & C. Best (eds.), *Laboratory phonology 8*. Berlin: Mouton de Gruyter. 51-79.
- Gafos, A. & S. Benus (2006). Dynamics of Phonological Cognition. *Cognitive Science* 30: 905-943.
- Ganong, E. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance* 6: 110-125.
- Garner, W. (1970). The stimulus in information processing. *American Psychologist* 25, 4: 350-358.
- Garner, W. (1974). *The processing of information and structure*. Potomac, MD: Lawrence Erlbaum.
- Garner, W. (1983). Asymmetric interaction of stimulus dimensions in perceptual information processing. In T. Tighe & B. Shepp (eds.), *Perception, cognition, and development: Interactional analysis*. Hillsdale, NJ: Lawrence Erlbaum. 1-38.
- Garner, W. & G. Felfoldy (1970). Integrality of stimulus dimensions in various types of information processing. *Cognitive Psychology* 1, 3: 225-241.
- Gaskell, M., P. Quinlan, J. Tamminen & A. Cleland (2008). The nature of phoneme representation in spoken word recognition. *Journal of Experimental Psychology: General* 137, 2: 282-302.
- Gay, T. (1974). A cinefluorographic study of vowel production. *Journal of Phonetics* 2: 255-266.
- Gay, T. (1977). Articulatory movements in VCV sequences. *Journal of the Acoustical Society of America* 62: 183-193.

- Gendrot, C. & M. Adda-Decker (2007). Impact of duration and vowel inventory size on formant values of oral vowels: An automated formant analysis from eight languages. *International Congress of Phonetic Sciences XVI*. Saarbrücken. 1417-1420.
- Gerfen, C. & K. Hall (2001). *Coda aspiration and incomplete neutralization in Eastern Andalusian Spanish*. Manuscript, University of North Carolina at Chapel Hill.
- Gernsbacher, M. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness and polysemy. *Journal of Experimental Psychology: General* 113: 256-281.
- Ghitza, O. & J. Goldstein (1983). JNDs for the spectral envelope parameters in natural speech. In R. Klinke & R. Hartmann (eds), *Hearing—Physiological Bases and Psychophysics*. Berlin: Springer-Verlag. 352-358.
- Giannini, A. & U. Cinque (1978). Phonetic status and phonemic function of the final devoiced stops in Polish. *Speech Laboratory Report*. Napoli: Istituto Universitario Orientale.
- Giard M., J. Lavikainen, K. Reinikainen, F. Perrin, O. Bertrand, J. Pernier & R. Näätänen (1995). Separate representation of stimulus frequency, intensity, and duration in auditory sensory memory. *Journal of Cognitive Neuroscience* 7, 2: 133-143.
- Giegerich, H. (1999). *Lexical strata in English: Morphological causes, phonological effects*. Cambridge: Cambridge University Press.
- Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues*. Hillsdale, NJ: Erlbaum. 311-339.
- Gigerenzer, G. & U. Hoffrage (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review* 102: 684-704.
- Gigerenzer G., S. Krauss & O. Vitouch (2004). The null ritual: What you always wanted to know about significance testing but were afraid to ask. In D. Kaplan (ed.), *The Sage handbook of quantitative methodology for the social sciences*. Thousand Oaks, CA: Sage Publications. 391-408.
- Gigerenzer, G. & D. Murray (1987). *Cognition as intuitive statistics*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Gliner, J., N. Leech & G. Morgan (2002). Problems with null hypothesis significance testing (NHST): What do the text books say? *The Journal of Experimental Education* 71, 1: 83-92.
- Gobl, C. & A. Ní Chasaide (1999). Laryngeal coarticulation: Section B: Voice source variation in the vowel as a function of consonantal contrast. In W. Hardcastle & N. Hewlett (eds.), *Coarticulation: Theory, data and techniques*. Cambridge: Cambridge University Press. 122-143.
- Godfrey, K. (1985). Statistics in practice: Comparing the means of several groups. *The New England Journal of Medicine* 313, 23: 1450-1456.
- Gold, D. (1969). Statistical tests and substantive significance. *The American Sociologist* 4: 42-46.
- Goldinger, S. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22: 1166-1183.

- Goldinger, S. (1997). Words and voices: Perception and production in an episodic lexicon. In K. Johnson, & J. Mullennix (eds.), *Talker variability in speech processing*. San Diego: Academic Press. 33–66.
- Goldinger, S. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review* 105: 251-279.
- Goldrick, M. (2001). Turbid output representations and the unity of opacity. In M. Hirotani, A. Coetzee, N. Hall & J-Y. Kim (eds.), *Proceedings of the Northeast Linguistics Society (NELS) 30*. Amherst, MA: GLSA. 231-245.
- Goldrick, M. & P. Smolensky (1999). Opacity, turbid representations, and output-based explanations. Paper presented at *the Workshop on the Lexicon in Phonetics and Phonology*. 11-13 June 1999, Alberta, Edmonton: University of Alberta.
- Gomes H., W. Ritter W & H. Vaughan (1995). The nature of preattentive storage in the auditory system. *Journal of Cognitive Neuroscience* 7: 81–94.
- Gordon, B. (1985). Subjective frequency and lexical decision latency function: Implications for mechanism of lexical access. *Journal of Memory and Language* 24, 6: 631-645.
- Gordon, M. (2001). Syncope induced metrical opacity as a weight effect. In K. Megerdooomian & L. Bar-El (eds.), *Proceedings of the 20th West Coast Conference on Formal Linguistics*: 206-219.
- Gordon, M. (2002). Investigating chain shifts and mergers. In J. Chambers, P. Trudgill, & N. Schilling-Estes (eds.), *The handbook of language variation and change*. Oxford: Blackwell. 244–266.
- Gouskova, M. & N. Hall (2007). Levantine Arabic epenthesis: Phonetics, phonology and learning. Poster presented at *the Workshop on Variation, Gradience and Frequency in Phonology*. 6-8 July 2007, Stanford, CA: Stanford University.
- Gouskova, M. & N. Hall (2009). Acoustics of epenthetic vowels in Lebanese Arabic. In S. Parker (ed.), *Phonological argumentation: Essays on evidence and motivation*. London: Equinox Publishing. 203-225.
- Granaas, M. (1998). Model fitting: A better approach. *American Psychologist* 53, 7: 800-810.
- Green, D. & J. Swets (1974). *Signal detection theory and psychophysics*. 2nd edition. New York: Wiley.
- Greenberg, J. (1950). The Patterning of root morphemes in Semitic. *Word* 6:162-181.
- Grier, J. (1971). Nonparametric indexes for sensitivity and bias: Computing formulas. *Psychological Bulletin* 75, 6: 424-429.
- Grosvald, M. (2009). *Long-distance coarticulation: A production and perception study of English and American Sign Language*. PhD thesis, University of California.
- Guenther, F. & M. Gjaja (1996). The perceptual-magnet effect as an emergent feature of neural map formation. *Journal of the Acoustical Society of America* 100:1111-1121.
- Gurevich, N. (2004). *Lenition and contrast: The functional consequences of certain phonologically conditioned sound changes*. New York: Routledge.
- Hagen, R. (1997). In praise of the null hypothesis statistical test. *American Psychologist* 52: 15-24.

- Hagstrom, P. (1997). Contextual metrical invisibility. In B. Bruening, Y. Kang & M. McGinnis (eds.), *PF: Papers at the interface. MIT Working Papers in Linguistics*. Cambridge, MA. 113-181.
- Haj Yusef, N. (1992). *Duration and some other phonetic aspects: An acoustic investigation of the Aleppo dialect of Syrian Arabic*. PhD thesis, University of Nottingham.
- Hall, D. (2007). *The role and representation of contrast in phonological theory*. PhD thesis, University of Toronto.
- Hall, K. (2008). *Testing an exemplar-based model of contrast and allophony against evidence from second language acquisition*. Manuscript, The Ohio State University.
- Hall, K. (2009). *A probabilistic model of phonological relationships from contrast to allophony*. PhD thesis, The Ohio State University.
- Hall, N. (2003). *Gestures and segments: Vowel intrusion as overlap*. PhD thesis, University of Massachusetts.
- Hall, N. (2006). Cross-linguistic patterns of vowel intrusion. *Phonology* 23: 387-429.
- Halle, M. (1995). Feature geometry and feature spreading. *Linguistic Inquiry* 26: 1-46.
- Halle, M. (2002). *From memory to speech and back: Papers on phonetics and phonology 1954-2002*. The Hague: Mouton de Gruyter.
- Haller, H. & S. Krauss (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research—MPR-online* 7, 1: 1-20. Retrieved 29 May 2009 from <http://www.dgps.de/fachgruppen/methoden/mpr-online/issue16/index.html>.
- Hall-Lew, L. (2009). *Ethnicity and phonetic variation in a San Francisco neighbourhood*. PhD thesis, Stanford University.
- Ham, W. (2001). *Phonetic and phonological aspects of geminate timing*. New York: Routledge.
- Hansson, G. (2008). Effects of contrast recoverability on the typology of harmony systems. In P. Avery, B. Dresher & K. Rice (eds.), *Contrast in phonology: Perception and acquisition*. Berlin: Mouton de Gruyter. 115-141.
- Harlow, L., S. Mulaik & J. Steiger (eds.) (1997). *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum Associates.
- Harris, J. (1997). Licensing inheritance: An integrated theory of neutralisation. *Phonology* 14: 315-370.
- Harris, J. & G. Lindsey (1995). The elements of phonological representation. In J. Durand & F. Katamba (eds.), *Frontiers of phonology: Atoms, structures, derivations*. London: Longman. 34-79.
- Harris, J. & G. Lindsey (2000). Vowel patterns in mind and sound. In N. Burton-Roberts, P. Carr & G. Docherty (eds.), *Phonological knowledge: Conceptual and empirical issues*. Oxford: Oxford University Press. 185-205.
- Harris, M. & N. Umeda (1987). Difference limens for fundamental frequency contours in sentences. *Journal of the Acoustical Society of America* 81, 4: 1139-1145.
- 't Hart, J. (1981). Differential sensitivity to pitch distance, particularly in speech. *Journal of the Acoustical Society of America* 69: 811-821.
- Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics* 31: 373-405.

- Hawkins, S. & N. Nguyen (2003). Effects on word recognition of syllable-onset cues to syllable-coda voicing. In J. Local, R. Ogden, & R. Temple (eds.), *Papers in laboratory phonology VI: Phonetic interpretation*. Cambridge: Cambridge University Press. 38-57.
- Hawkins, S. & N. Nguyen (2004). Influence of syllable-coda voicing on the acoustic properties of syllable-onset /l/ in English. *Journal of Phonetics* 32: 199-231.
- Hawkins, S. & A. Slater (1994). Spread of CV and V-to-V coarticulation in British English: Implications for the intelligibility of synthetic speech. *Proceedings of the 3rd International Conference on Spoken Language Processing (ICSLP-94)*. 57-60.
- Hay, J., P. Warren & K. Drager (2006). Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics* 34: 458-484.
- Hays, W. (1963). *Statistics*. New York: Harcourt Brace College Publishers.
- Henry, F. (1948). Discrimination of the duration of a sound. *Journal of Experimental Psychology* 38, 6: 734-743.
- Herd, W., A. Jongman & J. Sereno (2010). An acoustic and perceptual analysis of /t/ and /d/ flaps in American English. *Journal of Phonetics* 38, 4: 504-516.
- Herrick, D. (2004). Neutralisation is complete in Catalan vowel reduction. *Proceedings of the LSK 2004 International Conference*, Linguistic Society of Korea.
- Herzallah, R. (1990). *Aspects of Palestinian phonology: A non-linear approach*. PhD thesis, Cornell University.
- van Hessen, A. & M. Schouten (1992). Modelling phoneme perception II: A model of stop consonant discrimination. *Journal of the Acoustical Society of America* 92: 1856-1868.
- Hirai, Y. (1989). Some remarks on class interval of histograms. *Okayama University Bulletin of School of Education* 82, 1: 113-117.
- Hoaglin, D. (1983). Letter values: A set of selected order statistics. In D. Hoaglin, F. Mosteller & J. Tukey (eds.), *Understanding robust and exploratory data analysis*. New York: Wiley. 33-57.
- Hodges, J. & E. Lehmann (1954). Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society B16*: 261-268.
- Hodgson, P. & J. Miller (1996). Internal structure of phonetic categories: Evidence for within-category trading relations. *Journal of the Acoustical Society of America* 100, 1: 565-576.
- Houston, D. & P. Jusczyk (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance* 26, 5: 1570-1582.
- Howell, D. (2002). *Statistical methods for psychology*. 5th edition. Pacific Grove: Wadsworth Group.
- Howitt, D. & D. Cramer (2000). *An introduction to statistics in psychology: A complete guide for students*. 2nd edition. Harlow: Prentice-Hall.
- Hox, J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Hubbard, R. (1995). The earth is highly significantly round ($p < .0001$). *American Psychologist* 50: 1098.

- Hubbard, R. & J. Armstrong (2006). Why we don't really know what statistical significance means: Implications for educators. *Journal of Marketing Education* 28:114-120.
- Hubbard, R. & M. Bayrri (2003). Confusion over measures of evidence (p's) versus errors (α 's) in classical statistical testing. *The American Statistician* 57, 3: 171-182.
- Hubbard, R. & R. Lindsay (2008). Why p values are not a useful measure of evidence in statistical significance testing. *Theory and Psychology* 18, 1: 69-88.
- Huberty, C. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. *Journal of Experimental Education* 61: 317-333.
- Hudson, T., G. de Jong, K. McDougall, P. Harrison & F. Nolan (2007). F0 statistics for 100 young male speakers of Standard Southern British English. In J. Trouvain & W. Barry (eds.), *Proceedings of the 16th International Congress of Phonetic Sciences*, 6-10 August 2007, Saarbrücken. 1809-1812.
- Huff, C. (1980). Voicing and flap neutralization in New York City English. *Research in Phonetics* 1: 233-356.
- Huggins, A. (1972). Just noticeable differences for segment duration in natural speech. *Journal of the Acoustical Society of America* 51, 4B:1270-1278.
- Hunnicutt, S. (1985). Intelligibility vs. redundancy—conditions of dependency. *Language and Speech* 28: 47-56.
- Hunter, J. (1997). Needed: A ban on the significance tests. *Psychological Science* 8: 3-7.
- Huotilainen, M., R. Ilmoniemi, J. Lavikainen, H. Tiitinen, K. Alho, J. Sinkkonen, J. Knuutila & R. Näätänen (1993). Interaction between representations of different features of auditory sensory memory. *NeuroReport* 4: 1279-1281.
- Ibrahim, R., Z. Eviatar & J. Aharon-Peretz (2002). The characteristics of Arabic orthography slow its processing. *Neuropsychology* 16, 3: 322-326.
- Inouchi, M., M. Kubota, P. Ferrari & T. Roberts (2003). Magnetic mismatch fields elicited by vowels duration and pitch changes in Japanese word in humans: Comparison between native- and non-speakers of Japanese. *Neuroscience Letters* 353, 3: 165-168.
- Isačenko, A. & H. Schädlich (1970). *A model of standard German intonation*. The Hague: Mouton.
- Itô, J. & A. Mester (1999). The phonological lexicon. In N. Tsujimura (ed.), *The handbook of Japanese linguistics*. 62-100. Oxford: Blackwell.
- Itô, J. & A. Mester (2003). On the sources of opacity in OT: Coda processes in German. In: C. Féry & R. van de Vijver (eds.), *The Syllable in Optimality Theory*. Cambridge: Cambridge University Press. 271-303. [Available from Rutgers Optimality Archive, ROA-347].
- Jakobson, R., G. Fant & M. Halle (1963). *Preliminaries to speech analysis: The distinctive features and their correlates*. Cambridge, MA: MIT Press.
- Jansen, W. (2004). *Laryngeal contrast and phonetic voicing: A laboratory phonology approach to English, Hungarian, and Dutch*. PhD thesis, University of Groningen.
- Jassem, W. & L. Richter (1989). Neutralization of voicing in Polish obstruents. *Journal of Phonetics* 17: 317-325.

- Jenkins, S. (2002). Data pooling and type I errors: A comment on Leger & Didrichsons. *Animal Behaviour* 63: F9-F11.
- Johnson, E. & P. Jusczyk (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language* 44: 548-567.
- Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. Mullennix (eds.), *Talker variability in speech processing*: 145-165. San Diego: Academic Press.
- Johnson, K. (2001). Spoken language variability: Implications for modeling speech perception. In R. Smits, J. Kingston, T. Nearey & R. Zondervan (eds.), *Proceedings of the Workshop on Speech Recognition as Pattern Classification (SPRAAC)*. 11-13 July 2001, Nijmegen: Max Planck Institute for Psycholinguistics.
- Johnson, K. (2005). Speaker normalization in speech perception. In D. Pisoni & R. Remez (eds), *The Handbook of Speech Perception*. Oxford: Blackwell. 363-389.
- Johnson, K. (2007). Decisions and mechanisms in exemplar-based phonology. In M. Solé, P. Beddor & M. Ohala (eds.), *Experimental Approaches to Phonology*. Oxford: Oxford University Press. 25-40.
- Johnson, K. & J. Mullennix (eds.) (1997). *Talker variability in speech processing*. San Diego: Academic Press.
- Johnson, K., E. Strand & M. D'Imperio (1999). Auditory visual integration of talker gender in vowel perception. *Journal of Phonetics* 27: 359-384.
- Johnson, N. (1976). A note on the use of A' as a measure of sensitivity. *Journal of Experimental Child Psychology* 22: 530-531.
- Jones, L. & J. Tukey (2000). A sensible formulation of the significance test. *Psychological Methods* 5, 4: 411-414.
- Jongman, A. (2004). Phonological and phonetic representations: The case of neutralization. In A. Agwuele, W. Warren & S-H. Park (eds.), *Proceedings of the 2003 Texas Linguistics Society Conference*. Somerville, MA: Cascadilla Proceedings Project. 9-16.
- Kager, R. (2008). Lexical irregularity and the typology of contrast. In K. Hanson and S. Inkelas (eds.), *The nature of the word: Studies in honor of Paul Kiparsky*. Cambridge, MA: MIT Press. 397-432.
- Kainada, E. (2006). Segmental and suprasegmental cues to prosodic boundaries. In *Proceedings of the 15th Linguistics and English Language Postgraduate Conference*. 3 March 2006, Edinburgh: University of Edinburgh.
- Kaisse, E. (1985). *Connected speech: The interaction between syntax and phonology*. Orlando, Florida: Academic Press.
- Kamas, E., L. Reder & M. Ayers (1996). Partial matching in the Moses Illusion: Response bias not sensitivity. *Memory and Cognition* 24: 687-699.
- Kang, Y. (to appear). Loanword phonology. In M. van Oostendorp, C. Ewen, E. Hume & K. Rice (eds.), *The Blackwell Companion to Phonology*. Malden: Wiley-Blackwell.
- Kareev, Y., S. Arnon & R. Horwitz-Zeliger (2002). On the misperception of variability. *Journal of Experimental Psychology: General* 131, 2: 287-297.
- Katseff, S. & J. Houde (2008). Partial compensation in speech adaptation. *2008 Annual Report of the UC Berkeley Phonology Lab*: 444-461.

- Kaukoranta, E., M. Sams, R. Hari, M. Hämäläinen & R. Näätänen (1989). Reactions of human auditory cortex to changes in tone duration. *Hearing Research* 41: 15-22.
- Kazanina, N., C. Phillips & W. Idsardi (2006). The influence of meaning on the perception of speech sounds. *Proceedings of the National Academy of Sciences* 103, 30: 11381-11386.
- Keating, J. & D. Scott (1999). A primer on density estimation for the great homerun race of 1998. *Stats* 25: 16-22.
- Keating, P. (1985). Universal phonetics and the organization of grammars. In V. Fromkin (ed.), *Phonetic Linguistics*. New York: Academic Press. 115-32.
- Kenny, D. & C. Judd (1986). Consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin* 99, 3: 422-431.
- Kenstowicz, M. (1994). *Phonology in generative grammar*. Cambridge, MA: Blackwell Publishers.
- Kenstowicz, M. (1996). Base-identity and uniform exponence: Alternatives to cyclicity. In J. Durand & B. Laks (eds.), *Current trends in phonology: Models and methods*. Paris-X and Salford: University of Salford Publications. 363-393.
- Kenstowicz, M. (1997). Quality-sensitive stress. *Rivista di Linguistica* 9: 157-188. Reprinted in J. McCarthy (ed.) *Optimality Theory in phonology: A reader*. Oxford: Blackwell. 191-201.
- Kewley-Port, D. (1995). Thresholds for formant-frequency discrimination of vowels in consonantal context. *Journal of the Acoustical Society of America* 97: 3139-3146.
- Kewley-Port, D. (2001). Vowel formant discrimination II: Effects of stimulus uncertainty, consonantal context and training. *Journal of the Acoustical Society of America* 110: 2141-2155.
- Kewley-Port, D. & A. Neel (2006). Perception of dynamic properties of speech: Peripheral and central processes. In S. Greenberg & W. Ainsworth (eds.), *Listening to speech: An auditory perspective*. London: Lawrence Erlbaum. 49-61.
- Kewley-Port, D. & C. Watson (1994). Formant-frequency discrimination for isolated English vowels. *Journal of the Acoustical Society of America* 95: 485-496.
- Kewley-Port, D. & Y. Zheng (1999). Vowel formant discrimination: Towards more ordinary listening conditions. *Journal of the Acoustical Society of America* 106: 2945-2958.
- Kim, H., & A. Jongman (1996). Acoustic and perceptual evidence for complete neutralization of manner of articulation in Korean. *Journal of Phonetics* 24: 295-312.
- Kiparsky, P. (1971). Historical linguistics. In W. O. Dingwall (ed.), *A survey of linguistic science*. College Park: University of Maryland Linguistics Program. 576-642.
- Kiparsky, P. (1973). Abstractness, opacity, and global rules. In O. Fujimura (ed.), *Three dimensions in linguistic theory*. Tokyo: TEC. 57-86.
- Kirchner, R. (1996). Synchronic chain shifts in Optimality Theory. *Linguistic Inquiry* 27: 341-350.
- Kirchner, R. (1997). Contrastiveness and faithfulness. *Phonology* 14: 83-111.
- Kirchner, R. (1998). *An effort-based approach to consonant lenition*. PhD thesis, UCLA.
- Kirk, R. (1968). *Experimental design: Procedures for the behavioral sciences*. Pacific Grove: Brooks/Cole.

- Kirk, R. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement* 5: 746-759.
- Kirk, R. (2007). Effect magnitude: A different focus. *Journal of Statistical Planning and Inference* 137: 1634-1646.
- Kitto, C. & P. de Lacy (1999). Correspondence and epenthetic quality. In C. Kitto & C. Smallwood (eds.), *Toronto Working Papers in Linguistics*: 181-200.
- Klatt, D. (1973). Discrimination of fundamental frequency contours in synthetic speech: Implications for models of pitch perception. *Journal of the Acoustical Society of America* 53, 1: 8-16.
- Klatt, D. (1975). Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics* 3: 129-140.
- Klatt, D. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America* 59, 5: 1208-1221.
- Klatt, D. (1986). The problem of variability in speech recognition and in models of speech perception. In J. Perkell & D. Klatt (eds.), *Invariance and variability in speech processes*. Hillsdale, NJ: Lawrence Erlbaum Associates. 300-319.
- Klatt, D. (1992). Review of selected models of speech perception. In W. Marslen-Wilson (ed.), *Lexical representation and process*. 2nd edition. Cambridge, MA: MIT Press. 169-226.
- Klatt, D & W. Cooper (1975). Perception of segment duration in sentence contexts. In A. Cohen & S. Nooteboom (eds.), *Structure and process in speech perception*. Berlin: Springer-Verlag. 69-89.
- Klein, S. (2001). Measuring, estimating, and understanding the psychometric function: A commentary. *Perception and Psychophysics* 63, 8: 1421-1455.
- Kochanski, G. (2006). Prosody beyond fundamental frequency. In S. Sudhoff, D. Lenertová, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter & J. Schließer (eds.), *Methods in empirical prosody research*. Berlin: Walter de Gruyter. 89-122.
- Kochanski, G., J. Coleman & B. Rosner (2005). Loudness predicts prominence: Fundamental frequency lends little. *Journal of the Acoustical Society of America* 118, 2: 1038-1054.
- Kopkallı, H. (1993). *A phonetic and phonological analysis of final devoicing in Turkish*. PhD thesis, The University of Michigan.
- Kraus, N., T. McGee, T. Carrell & A. Sharma (1995). Neurophysiologic bases of speech discrimination. *Ear and Hearing* 16: 19-37.
- Kuhl, P. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics* 50, 2: 93-107.
- Kuhl, P. & P. Iverson. (1995). Linguistic experience and the perceptual magnet effect. In W. Strange (ed.), *Speech perception and linguistic experience: Issues in cross-language research*. Timonium, MD: York Press. 121-154.
- Kühnert, B. & F. Nolan (1999). The origin of coarticulation. In W. Hardcastle & N. Hewlett (eds.), *Coarticulation: Theory, data and techniques*. Cambridge: Cambridge University Press. 7-30.

- Labov, W. (1986). Sources of inherent variations in the speech process. In J. Perkell & D. Klatt (eds.), *Invariance and variability in speech processes*. Hillsdale, NJ: Lawrence Erlbaum Associates. 402-425.
- Labov, W. (1994). *Principles of linguistic change I: Internal factors*. Oxford: Blackwell.
- Labov, W., M. Yaeger & R. Steiner (1972). *A Quantitative study of sound change in progress*. Report on National Science Foundation Project no. GS-3287. Philadelphia: US Regional Survey.
- Lacerda, F. (1997). *Distributed memory representations generate the perceptual-magnet effect*. Manuscript, Institute of Linguistics, Stockholm University.
- Lachs, L., K. McMichael & D. Pisoni (2003). Speech perception and implicit memory: evidence for detailed episodic encoding. In J. Bowers & C. Marsolek (eds.), *Rethinking implicit memory*. Oxford: Oxford University Press. 215-235.
- Ladd, D. (2006). "Distinctive phones" in surface representation. In L. Goldstein, D. Whalen & C. Best (eds.), *Laboratory phonology 8*. Berlin: Mouton de Gruyter. 3-26.
- Ladefoged, P. (1967). Linguistic phonetics. *Working Papers in Phonetics* 6. Department of Linguistics, UCLA.
- Ladefoged, P. (1993). *A Course in Phonetics*. 3rd edition. Fort Worth: Harcourt Brace College.
- Lahiri, A. & J. Hankamer (1988). The timing of geminate consonants. *Journal of Phonetics* 16: 327-338.
- Lahiri, A., & W. Marslen-Wilson (1991). The mental representation of lexical form: a phonological approach to the recognition lexicon. *Cognition* 38, 3: 245-294.
- Lahiri, A. & H. Reetz (2002). Underspecified recognition. In C. Gussenhoven, N. Warner & T. Rietveld (eds.), *Laboratory phonology 7*. Berlin: Mouton. 637-676.
- Lang, A., T. Nyeke, M. Ek, O. Aaltonen, Y. Raimo & R. Näätänen (1990). Pitch discrimination performance and auditory event-related potentials. In C. Brunia, A. Gaillard & A. Kok (eds.), *Psychophysiological brain research I*. Tilburg: Tilburg University Press. 294-298.
- Lann, A. & R. Falk (2003). What are the clues for intuitive assessment of variability? In C. Lee (ed.), *Proceedings of the 3rd International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-3)*. Mount Pleasant, MI: Central Michigan University.
- Lapid, E., R. Ulrich & T. Rammsayer (2008). On estimating the difference limen in duration discrimination tasks: A comparison of the 2AFC and the reminder task. *Perception and Psychophysics* 70, 2: 291-305.
- Lathrop, R. (1967). Perceived variability. *Journal of Experimental Psychology* 73, 4: 498-502.
- Laver, J. (1994) *Principles of phonetics*. Cambridge: Cambridge University Press.
- Lavoie, L. (2001). *Consonant strength: Phonological patterns and phonetic manifestations*. New York: Garland.
- Lavoie, L. (2002). Subphonemic and suballophonic consonant variation: The role of the phoneme inventory. *ZAS Papers in Linguistics* 28: 39 -54.
- Lee, S. (1991). The duration and perception of English epenthetic and underlying stops. *Journal of the Acoustical Society of America* 89, 4B: 1999.

- Lee, Y. (2001). The noun-verb asymmetry in Korean phonology. *Studies in Phonetics, Phonology and Morphology* 7, 2: 375-397.
- van der Leeden, R. (1998). Multilevel analysis of repeated measures data. *Quality and Quantity* 32: 15-19.
- Leek, M. (2001). Adaptive procedures in psychophysical research. *Perception and Psychophysics* 63, 8: 1279-1292.
- Leger, D. & I. Didrichsons (1994). An assessment of data pooling and some alternatives. *Animal Behaviour* 48: 823-832.
- Lehiste, I. (1970). *Suprasegmentals*. Cambridge, MA: MIT Press.
- Lehiste, I. & G. Peterson (1959). Linguistic considerations in the study of speech intelligibility. *Journal of the Acoustical Society of America* 31, 3: 280-286.
- Levin, J. (1987). Between epenthetic and excrescent vowels. In M. Crowhurst (ed.), *Proceedings of the sixth West Coast Conference on Formal Linguistics*: 187-201.
- Levin, J. & J. Miller (1996). Broadband neural encoding in the cricket cercal sensory system enhanced by stochastic resonance. *Nature* 380, 6570: 165-168.
- Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech* 6: 172-187.
- Lindblom, B. (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America* 35, 11: 1773-1781.
- Lindblom, B. (1986). Phonetic universals in vowel systems. In J. Ohala & J. Jaeger (eds.), *Experimental phonology*. Orlando: Academic Press. 13-44.
- Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H&H theory. In W. Hardcastle & A. Marchal (eds.), *Speech production and speech modeling*. Dordrecht: Kluwer Academic Publishers. 403-439.
- Liphola, M. (2001). *Aspects of phonology and morphology of Shimakonde*. PhD thesis, The Ohio State University.
- Lisker, L. & A. Abramson (1967). Some effects of context on voice onset time in English stops. *Language and Speech* 10: 1-28.
- Lively, S. (1993). An examination of the perceptual magnet effect. *Journal of the Acoustical Society of America* 93, 4: 2423.
- Lively, S., J. Logan & D. Pisoni (1993). Training Japanese listeners to identify English /r/ and /l/: II. The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America* 94: 1242-1255.
- Lloret, M-R. & J. Jiménez (2006). Prominence-driven epenthesis in Alguerese Catalan. Paper presented at the *Third Old World Conference on Phonology (POCP-3)*. 17-19 January 2006, Budapest.
- Loakes, D. (2006). Variation in long-term fundamental frequency: Measurements from vocalic segments in twins' speech. *Proceedings of the 11th Australian International Conference on Speech Science & Technology*, University of Auckland, New Zealand. December 6-8 2006, Auckland. 205-210. <<http://www.assta.org/sst/2006/>>.
- Locker, L., L. Hoffman & J. Bovaird (2007). On the use of multilevel modeling as an alternative to items analysis in psycholinguistic research. *Behavior Research Methods* 39, 4: 723-730.

- Lombardi, L. (1994). *Laryngeal features and laryngeal neutralization*. New York: Garland.
- Lombardi, L. (2002). Markedness and the typology of epenthetic vowels. Paper presented at *the Workshop in the Linguistics and Phonetics 2002 (LP2002) Conference*. 2-6 September 2002, Chiba, Japan: Meikai University. <http://www.adr.jp/~ad31175/lp2002/lp2002main.htm>.
- Lotto, A., K. Kluender & L. Holt (1998). Depolarizing the perceptual magnet effects. *Journal of the Acoustical Society of America* 103, 6: 3648-3655.
- Lotto, A., K. Kluender & L. Holt (2000). Effects of language experience on organization of vowel sounds. In M. Broe & J. Pierrehumbert (eds.), *Laboratory phonology V: Language acquisition and the lexicon*. Cambridge: Cambridge University Press. 219-228.
- Lovie, P. & A. Lovie (1976). Teaching intuitive statistics I: Estimating means and variances. *International Journal of Mathematical Education in Science and Technology* 7, 1: 29-39.
- Łubowicz, A. (2002). Derived environment effects in Optimality Theory. *Lingua* 112: 243-280.
- Łubowicz, A. (2003). *Contrast preservation in phonological mappings*. PhD thesis, MIT.
- Luce, P. & C. McLennan (2005). Spoken word recognition: The challenge of variation. In D. Pisoni & R. Remez (eds.), *The handbook of speech perception*. Malden, MA: Blackwell. 591-609.
- Machlis, L., W. Dodd & J. Fentress (1985). The pooling fallacy: Problems arising when individuals contribute more than one observation to the data set. *Zeitschrift für Tierpsychologie* 68, 3: 201-214.
- Macmillan, N. (2001). Threshold estimation: The state of the art. *Perception and Psychophysics* 63, 8: 1277-1278.
- Macmillan, N. & C. Creelman (1990). Response bias: Characteristics of detection theory, threshold theory, and "nonparametric" measures. *Psychological Bulletin* 107: 401-413.
- Macmillan, N. & C. Creelman (2005). *Detection theory: A user's guide*. 2nd edition. New York: Psychology Press.
- Magen, H. (1997). The extent of vowel-to-vowel coarticulation in English. *Journal of Phonetics* 25: 187-205.
- Manaster Ramer, A. (1996a). A letter from an incompletely neutral phonologist. *Journal of Phonetics* 24, 4: 477-489.
- Manaster Ramer, A. (1996b). Report on Alexis' dreams—bad as well as good. *Journal of Phonetics* 24, 4: 513-519.
- Manguson, J., R. Yamada, Y. Tohkura, D. Pisoni, S. Lively & A. Bradlow (1995). The role of talker variability in non-native phoneme training. *Proceedings of the 1995 Spring Meeting of the Acoustical Society of Japan*. 393-394.
- Maniwa, K. (2002). Acoustic and perceptual evidence of complete neutralization of word-final tonal specification in Japanese. In S. Stowers & N. Poell (eds.), *Kansas Working Papers in Linguistics* 26: 93-112.
- Manuel, S. (1990). The role of contrast in limiting vowel-to-vowel coarticulation in different languages. *Journal of the Acoustical Society of America* 88: 1286-1298.

- Manuel, S. (1999). Cross-language studies: Relating language-particular coarticulation patterns to other language-particular facts. In W. Hardcastle & N. Hewlett (eds.), *Coarticulation: Theory, data and techniques*. Cambridge: Cambridge University Press. 179-198.
- Manuel, S. & R. Krakow (1984). Universal and language particular aspects of vowel-to-vowel coarticulation. *Haskins Laboratories Status Reports on Speech Research*: 69-78.
- Marsh, H., A. O'Mara & L. Malmberg (2008). Meta-analysis: A three-level multilevel meta-analysis. *ESRC RDI Meta-analysis Workshop*. Retrieved 20th February 2009 from <http://www.education.ox.ac.uk/research/researchgroup/self/training.php>.
- Marslen-Wilson, W. & P. Warren (1994). Levels of perceptual representation and process in lexical access: Words, phonemes, and features. *Psychological Review* 101, 4: 653-675.
- Mascaró, J. (1987). Underlying voicing recoverability of finally devoiced obstruents in Catalan. *Journal of Phonetics* 15: 183-186.
- Max, L. & P. Onghena (1999). Some issues in the statistical analysis of completely randomized and repeated measures designs for speech, language, and hearing research: Tutorial. *Journal of Speech, Language, and Hearing Research* 42: 261-270.
- McCarthy, J. (1986). OCP effects: Gemination and Antigemination. *Linguistic Inquiry* 7: 207-263.
- McCarthy, J. (1991). Semitic gutturals and distinctive features theory. In B. Comrie & M. Eid (eds.), *Perspectives on Arabic linguistics III: Papers from 3rd annual symposium on Arabic linguistics*. Amsterdam: John Benjamins. 63-91.
- McCarthy, J. (1994). The phonetics and phonology of Semitic pharyngeals. In P. Keating (ed.), *Papers in laboratory phonology III: Phonological structure and phonetic form*. Cambridge: Cambridge University Press.
- McCarthy, J. (1999). Sympathy and phonological opacity. *Phonology* 16: 331-399.
- McCarthy, J. (2003). Sympathy, cumulativity, and the Duke-of-York gambit. In C. Féry & R. van de Vijver (eds.), *The syllable in Optimality Theory*. Cambridge: Cambridge University Press. 23-76.
- McCarthy, J. (2006). Candidates and derivations in Optimality Theory. Talk presented at MIT. [Available at http://socrates.berkeley.edu/~ling298/roa_823.pdf].
- McCarthy, J. (2007a). *Hidden generalizations: Phonological opacity in Optimality Theory*. London: Equinox.
- McCarthy, J. (2007b). Where's opacity? Handout of a talk at *the 15th Manchester Phonology Meeting (MFM)*. 24-26 May 2007, Manchester: University of Manchester.
- McCarthy, J. & A. Prince (1995). Faithfulness and reduplicative identity. In J. Beckman, L. Walsh-Drckey & S. Urbanczyk (eds.), *University of Massachusetts Occasional Papers in Linguistics 18: Papers in Optimality Theory*. Amherst, MA: Graduate Linguistic Student Association.
- McCloskey, D. & S. Ziliak (1996). The standard error of regression. *Journal of Economic Literature* 34: 97-114.
- McClure, J. & H. Suen (1994). Interpretation of statistical significance testing: A matter of perspective. *Topics in Early Children Special Education* 14: 88-102.

- McDougall, K. (2004). Speaker-specific formant dynamics: an experiment on Australian English /a₁/. *International Journal of Speech, Language and the Law* 11, 1: 103-130.
- McDougall, K. (2006). Dynamic features of speech and the characterisation of speakers: towards a new approach using formant frequencies. *International Journal of Speech, Language and the Law* 13, 1: 89-126.
- McIntosh, A., N. Kovacevic & R. Itier (2008). Increased brain signal variability accompanies lower behavioral variability in development. *PLoS Computational Biology* 4, 7: e1000106.
- McMichael, K. (1999). Talker-specific encoding effects in recognition memory for sentences. *Research in Spoken Language Processing: Progress Report 23*. Indiana University. 166-198.
- McMurray, B., M. Tanenhaus & R. Aslin (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition* 86: B33-B42.
- McQueen, J., A. Cutler & D. Norris (2006). Phonological abstraction in the mental lexicon. *Cognitive Science* 30: 1113-1126.
- Meehl, P. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. Harlow, S. Mulaik & J. Steiger (eds.), *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum Associates. 393-426.
- Menon, R. (1993). Statistical significance testing should be discontinued in mathematics education research. *Mathematics Education Research Journal* 5, 1: 4-18.
- Mermelstein, P. (1978). Difference limens for formant frequencies of steady-state and consonant-bound vowels. *Journal of the Acoustical Society of America* 63, 2: 572-580.
- Meunier, C., R. Espesser & C. Frenck-Mestre (2006). Phonetic variability as a static/dynamic process in speech communication: a cross linguistic study. Paper presented at the 10th Conference on Laboratory Phonology (LabPhon 10). 30 June-2 July 2006, Paris: Université de Paris.
- Miller, J., C. Connine, T. Schermer & K. Kluender (1983). A possible auditory basis for internal structure of phonetic categories. *Journal of the Acoustical Society of America* 73, 6: 2124-2133.
- Miller, J. & L. Volaitis (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception and Psychophysics* 46, 6: 505-512.
- Milliken, G. & D. Johnson (2009). *Analysis of messy data, Vol. 1: Designed experiments*. New York: CRC Press.
- Milroy, J. & J. Harris (1980). When is a merger not a merger? The MEAT/MATE problem in a present-day English vernacular. *English World-Wide* 1: 199-210.
- Mitleb, F. (1981). Temporal correlates of 'voicing' and its neutralization in German. *Research in Phonetics* 2: 173-191.
- Mitleb, F. (1984). Voicing effect on vowel duration is not an absolute universal. *Journal of Phonetics* 12: 23-7.
- Mokros, J. & S. Russell (1995). Children's concept of average and representativeness. *Journal for Research in Mathematics Education* 26, 1: 20-39.

- Moon, S. & B. Lindblom (1994). Interaction between duration, context, and speaking style in English stressed vowels. *Journal of the Acoustical Society of America* 96, 1: 40-55.
- Moore, B. (1973). Frequency difference limens for short-duration tones. *Journal of the Acoustical Society of America* 54, 5: 610-619.
- Morais, J. & R. Kolinsky (1995). Perception and awareness in phonological processing: The case of the phoneme. In J. Mehler & S. Franck (eds.), *Cognition on cognition*. Amsterdam: Elsevier. 349-359.
- Morrison, D. & K. Henkel (1970). Significance tests in behavioral research: Pessimistic conclusions and beyond. In D. Morrison & K. Henkel (eds.), *The significance test controversy*. Chicago: Aldine. 305-311.
- Morrison, G. (2008). Perception of synthetic vowels by monolingual Canadian-English, Mexican-Spanish, and Peninsular-Spanish listeners. *Canadian Acoustics* 36, 4: 17-23.
- Morton, J. & W. Jassem (1965). Acoustic correlates of stress. *Language and Speech* 8: 159-181.
- Moulton, K. (2003). Deep allophones in the Old English laryngeal system. *Toronto Working Papers in Linguistics* 20: 157-173.
- Mullennix, J. & D. Pisoni (1990). Stimulus variability and processing dependencies in speech perception. *Perception and Psychophysics* 47, 4: 379-390.
- Mullennix, J., D. Pisoni & C. Martin (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America* 85: 365-378.
- Munson, B. & N. Solomon (2004). The effect of phonological neighbourhood density on vowel articulation. *Journal of Speech, Language, and Hearing Research* 47: 1048-1058.
- Myers, J. & J. Tsay (2008). Neutralization in Taiwan Southern Min Tone sandhi. In Y. Hsiao, H-C. Hsu, L-H. Wee & D-A. Ho (eds.), *Interfaces in Chinese phonology: Festschrift in honor of Matthew Y. Chen on his 70th birthday*. Taipei: Academia Sinica. 47-78.
- Myers, S. & B. Hansen (2007). The origin of vowel length neutralisation in final position: Evidence from Finnish speakers. *Natural Language and Linguistic Theory* 25: 157-193.
- Näätänen, R. (1992). *Attention and brain function*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Näätänen, R. (2001). The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent (MMNm). *Psychophysiology* 38: 1-21.
- Näätänen, R. & K. Alho (1995). Mismatch negativity—a unique measure of sensory processing in audition. *International Journal of Neuroscience* 80: 317-337.
- Näätänen, R., A. Lehtokoski, M. Lennes, M. Cheour, M. Huotilainen, A. Livonen, M. Vainio, P. Alku, R. Ilmoniemi, A. Luuk, J. Allik, J. Sinkkonen & K. Alho (1997). Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature* 385: 432-434.
- Näätänen, R., P. Paavilainen, T. Rinne & K. Alho (2007). The mismatch negativity (MMN) in basic research of central auditory processing: A review. *Clinical Neurophysiology* 118: 2544-2590.

- Nääätänen, R., E. Schröger, S. Karakas, M. Tervaniemi & P. Paavilainen (1993). Development of a memory trace for a complex sound in the human brain. *NeuroReport* 4: 503-506.
- Nääätänen, R. & I. Winkler (1999). The concept of auditory stimulus representation in cognitive neuroscience. *Psychological Bulletin* 125: 826-859.
- Nakagawa, S. & T. Foster (2004). The case against retrospective statistical power analyses with an introduction to power analysis. *Acta Ethologica* 7, 2: 103-108.
- Neath, I. & A. Surprenant (2003). *Human memory: An introduction to research, data, and theory*. 2nd edition. Australia: Thomson.
- Nespor, M. & I. Vogel (1986). *Prosodic phonology*. Dordrecht: Foris.
- Nevins, A. (2007). Book review of Jonathan Barnes' "Strength and weakness at the interface: Positional neutralization in phonetics and phonology". *Phonology* 24, 3: 461-469.
- Nguyen, N., S. Wauquier & B. Tuller (2009). The dynamical approach to speech perception: From fine phonetic detail to abstract phonological categories. In F. Pellegrino, E. Marsico, I. Chitoran & C. Coupé (eds.), *Approaches to phonological complexity*. Berlin: Walter de Gruyter. 191-218.
- Nickerson, R. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods* 5: 241-301.
- Nielsen, K. (2008). *Word-level and feature-level effects in phonetic imitation*. PhD thesis, University of California.
- Nishinuma, Y., A. Di Cristo & R. Espesser (1983). Loudness as a function of vowel duration in CV syllables. *Speech Communication* 2: 167-169.
- Nitko, A. (1983). *Educational tests and measurement: An introduction*. New York: Harcourt Brace Jovanovich.
- Nolan, F. (1983). *The phonetic bases of speaker recognition*. Cambridge: Cambridge University Press.
- Nolan, F. (1985). Idiosyncrasy in coarticulatory strategies. *Cambridge Papers in Phonetics and Experimental Linguistics* 4: N1-N9.
- Nolan, F. (1986). The implications of partial assimilation and incomplete neutralisation. *Cambridge Papers in Phonetics and Experimental Linguistics* 5: N1-N11.
- Nolan, F. (1992). The descriptive role of segments: Evidence from assimilation. In D. Docherty & D. Ladd (eds.), *Papers in laboratory phonology II: Gesture, segment, prosody*. Cambridge: Cambridge University Press. 261-280.
- Nolan, F., K. McDougall, G. de Jong & T. Hudson (2006). A forensic phonetic study of 'dynamic' sources of variability in speech: the DyViS project. *Proceedings of the 11th Australian International Conference on Speech Science and Technology*, University of Auckland, New Zealand. 6-8 December 2006, Auckland. 13-18. <http://www.assta.org/sst/2006/>
- Nosofsky, R. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General* 115: 39-57.
- Nunberg, G. (1980). A falsely reported merger in eighteenth-century English: A study in diachronic variation. In W. Labov (ed.), *Locating language in time and space*. New York: Academic Press. 221-250.

- Nycz, J. (2005). The dynamics of near merger in accommodation. *Proceedings of ConSOLE XIII, 2005*. 273-285.
- Nygaard, L. (2005). Perceptual integration of linguistic and non-linguistic properties of speech. In D. Pisoni & R. Remez (eds.), *The handbook of speech perception*. London: Blackwell. 390-413.
- Nygaard, L., M. Sommers & D. Pisoni (1994). Speech perception as a talker contingent process. *Psychological Science* 5: 42-46.
- Nygaard, L., M. Sommers & D. Pisoni (1995). Effects of stimulus variability on perception and representation of spoken words in memory. *Perception and Psychophysics* 57, 7: 989-1001.
- Ogasawara, N. (2007). *Processing of speech variability: Vowel reduction in Japanese*. PhD thesis, University of Arizona.
- Oh, E. (1998). Stress patterns of Bedouin Hijazi Arabic: An OT account. *Kansas Working Papers in Linguistics* 23. University of Kansas, Linguistics Graduate Student Association. 17-26.
- Ohala, J. (1983). The phonological end justifies any means. In S. Hattori & K. Inoue (eds.), *Proceedings of the XIIIth International Congress of Linguists*, 29 August–4 September, Tokyo. 232-243.
- Ohala, J. & D. Feder (1994). Listeners' normalization of vowel quality is influenced by 'restored' consonantal context. *Phonetica* 51: 111-118.
- Olive, J., A. Greenwood & J. Coleman (1993). *Acoustics of American English speech: A dynamic approach*. New York: Springer.
- van Oostendorp, M. (2008). Incomplete devoicing in formal phonology. *Lingua* 118, 9: 1362-1374.
- van Oostendorp, M. & A. Revithiadou (in progress). *Quasi-opacity and headed spans in Silly and Megisti Greek*. Manuscript, University of the Aegean and Meertens Institute.
- Orgun, C. (1996). Correspondence and identity constraints in two-level Optimality Theory. In J. Camecho, L. Choueiri & M. Watanabe (eds.), *Proceedings of the Fourteenth West Coast Conference on Formal Linguistics*. 399-413.
- Osborne, J. (2002). Notes on the use of data transformations. *Practical Assessment, Research and Evaluation* 8, 6. Retrieved 15 January 2009 from <http://PAREonline.net/getvn.asp?v=8&n=6>
- Padgett J. & M. Tabain (2005). Adaptive dispersion theory and phonological vowel reduction in Russian. *Phonetica* 62:14-54.
- Palmeri, T., S. Goldinger & D. Pisoni (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory and Cognition* 19, 2: 309-328.
- Paradis, C. & J-F. Prunet (eds.) (1991). *Phonetics and phonology II. The special status of coronals: Internal and external evidence*. New York: Academic Press.
- Pearce, M. (2007). ATR allophones or undershoot in Kera? In R. Breheny & N. Velegrakis (eds.), *UCL Working Papers in Linguistics* 19: 31-43.
- Pedhazur, E. & L. Schmelkin (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Erlbaum.

- Peng, S-h. (2000). Lexical versus 'phonological' representations of Mandarin sandhi tones. In B. Broe & J. Pierrehumbert (eds.), *Papers in laboratory phonology V: Acquisition and the lexicon*. Cambridge: Cambridge University Press. 152-167.
- Perkell, J. (1996). Properties of the tongue help to define vowel categories: Hypotheses based on physiologically oriented modelling. *Journal of Phonetics* 24: 3-22.
- Perkell, J. & D. Klatt (eds.) (1986). *Invariance and variability in speech processes*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Peterson, C. & L. Beach (1967). Man as an intuitive statistician. *Psychological Bulletin* 68, 1: 29-46.
- Peterson, G. & H. Barney (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America* 24: 175-184.
- Picton, T., C. Alain, L. Otten, W. Ritter & A. Achim (2000). Mismatch negativity: Different waters in the same river. *Audiology and Neuro-Otology* 5: 111-139.
- Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee & P. Hooper (eds.), *Frequency effects and the emergence of lexical structure*. Amsterdam: John Benjamins. 137-157.
- Pierrehumbert, J. (2002). Word-specific phonetics. In C. Gussenhoven & N. Warner (eds.), *Laboratory Phonology 7*. Berlin: Mouton de Gruyter. 101-139.
- Piggott, G. (1995). Epenthesis and syllable weight. *Natural Language and Linguistic Theory* 13: 283-326.
- Pingel, L. (1993). Variability—Does the standard deviation always measure it adequately? *Teaching Statistics* 15, 3: 70-71.
- Piroth, H. & P. Janker (2004). Speaker-dependent differences in voicing and devoicing of German obstruents. *Journal of Phonetics* 32: 81-109.
- Piroth, H., L. Schiefer, P. Janker & B. Johne (1991). Evidence for final devoicing in German? An experimental investigation. In *Proceedings of the 12th International Congress of the Phonetic Sciences*.
- Pisoni, D. (1997). Some thoughts on "normalization" in speech perception. In K. Johnson & J. Mullennix (eds.), *Talker variability in speech processing*. San Diego: Academic Press. 9-32.
- Pluymaekers, M., M. Ernestus & H. Baayen (2005). Lexical frequency and acoustic reduction in spoken Dutch. *Journal of the Acoustical Society of America* 118, 4: 2561-2569.
- Pols, L. (1999). Flexible, robust, and efficient human speech processing versus present-day speech technology. *Proceedings of the XIVth International Congress of Phonetic Sciences (ICPhS'99)*, San Francisco, CA. 9-16.
- Poon, P. & C. Mateer (1985). A study of VOT in Nepali stop consonants. *Phonetica* 42: 39-47.
- Port, R. (1977). *The influence of speaking tempo on the duration of stressed vowel and medial Stop in English trochee words*. Indiana University Linguistics Club, Bloomington.
- Port, R. (1996). The discreteness of phonetic elements and formal linguistics: Response to A. Manaster Ramer. *Journal of Phonetics* 24, 4: 491-511.
- Port, R. & P. Crawford (1989). Pragmatic effects on neutralization rules. *Journal of Phonetics* 16: 257-282.

- Port, R. & A. Leary (2005). Against formal phonology. *Language* 85: 927-964.
- Port, R. & M. O'Dell (1985). Neutralization of syllable-final voicing in German. *Journal of Phonetics* 13: 455-471.
- Portillo, P. (2002). Towards more natural synthetic speech. *Procesamiento del Lenguaje Natural* 29: 165-172.
- Pycha, A. (2006). A duration based solution to the problem of stress realization in Turkish. *UC Berkeley Phonology Lab Annual Report*: 141-151.
- Pye, S. (1986). Word-final devoicing of obstruents in Russian. *Cambridge Papers in Phonetics and Experimental Linguistics* 5: P1-P10.
- Quené, H. (2007). On the just noticeable difference for tempos in speech. *Journal of Phonetics* 35, 3: 353-362.
- Quené, H. & H. van den Bergh (2004). On multi-level modelling of data from repeated measures designs: A tutorial. *Speech Communication* 43: 103-121.
- Quené, H. & H. van den Bergh (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language* 59: 413-425.
- Raaijmakers, J. (2003). A further look at the "language-as-fixed-effect fallacy". *Canadian Journal of Experimental Psychology* 57, 3: 141-151.
- Raaijmakers, J., J. Schrijnemakers & F. Gremmen (1999). How to deal with "the language-as-fixed-effect fallacy": Common misconceptions and alternative solutions. *Journal of Memory and Language* 41: 416-426.
- Rallo Fabra, L. (2006). Can adult ESL learners discriminate the difference between L1 and L2 vowels? Evidence from native Catalan speakers learning English. *Processes and Process-Oriented in Foreign Language Teaching and Learning*: 1035-1042.
- Rasbash, J., C. Charlton, W. Browne, M. Healy & B. Cameron (2009a). *MLwiN Version 2.1*. Centre for Multilevel Modelling, University of Bristol.
- Rasbash, J., F. Steele, W. Browne & H. Goldstein (2009b). *A user's guide to MLwiN 2.10*. Centre for Multilevel Modelling, University of Bristol.
- Recasens, D. (1985). Coarticulatory patterns and degrees of coarticulatory resistance in Catalan CV sequences. *Language and Speech* 28: 97-114.
- Recasens, D. (1987). An acoustic analysis of V-to-C and V-to-V coarticulatory effects in Catalan and Spanish VCV sequences. *Journal of Phonetics* 15: 299-312.
- Recasens, D. (1989). Long range coarticulatory effects for tongue dorsum contact in VCVCV sequences. *Haskins Laboratories Status Report on Speech Research SR-99/100*: 19-37.
- Recasens, D. & A. Espinosa (2009). Dispersion and variability in Catalan five and six peripheral vowel systems. *Speech Communication* 51: 240-258.
- Remijsen, B. & L. Gilley (2008). Why are three-level vowel length systems rare? Insights from Dinka (Luanyjang dialect). *Journal of Phonetics* 36, 2: 318-344.
- Repp, B. & A. Liberman (1987). Phonetic category boundaries are flexible. In S. Harnad (ed.), *Categorical perception: The groundwork of cognition*. Cambridge: Cambridge University Press. 89-112.
- Repp, B. & H. Lin (1989). Acoustic properties and perception of stop consonant release transients. *Journal of the Acoustical Society of America* 85, 1: 379-396.

- Ridjanovic, M. (1986). On the struggle of underlying vowels for a voice in surface phonetic structure: Evidence from Serbo-Croatian. *Journal of the Acoustical Society of America* 79, S1: S26-S27.
- Ridouane, R. (2007). Gemination in Tashlhiyt Berber: An acoustic and articulatory study. *Journal of the International Phonetic Association* 37, 2: 119-142.
- Rietveld, T. & R. van Hout (2005). *Statistics in language research: Analysis of variance*. New York: Mouton de Gruyter.
- Robey, R. (2004). Reporting point and interval estimates of effect-size for planned contrasts: Fixed within effect analyses of variance. *Journal of Fluency Disorders* 29, 4: 307-341.
- Roman, G. & B. Pavard (1987). A comparative study: How we read Arabic and French. In J. K. O'Regan & A. Levy-Schoen (eds.), *Eye movements: From physiology to cognition*. Amsterdam: North Holland Elsevier. 431-440.
- van Rooy, B., D. Wissing & D. Paschall (2003). Demystifying incomplete neutralisation during final devoicing. *Southern African Linguistics and Applied Language Studies* 21: 49-66.
- Rose, P. (2002). *Forensic Speaker Identification*. London: Taylor & Francis.
- Rose, S. & L. King (2007). Speech error elicitation and co-occurrence restrictions in two Ethiopian Semitic languages. *Language and Speech* 50, 4: 451-504.
- Rosenberger, J. & M. Gasko (1983). Comparing location estimators: Trimmed means, median, and trimean. In D. Hoaglin, F. Mosteller & J. Tukey (eds.), *Understanding robust and exploratory data analysis*. New York: Wiley. 297-335.
- Rosenblith, W. & K. Stevens (1953). On the DL for frequency. *Journal of the Acoustical Society of America* 25, 5: 980-985.
- Rossi, M. & M. Chafcouloff (1972). Recherche sur le seuil différentiel de fréquence fondamentale dans la parole. *Travaux de L'institut de Phonétique d'aix* 1: 179-185.
- Rost, G. & B. McMurray (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science* 12, 2: 339-349.
- Rothman, K. (1990). No adjustments are needed for multiple comparisons. *Epidemiology* 1, 1: 43-46.
- Rudin, C. (1980). Phonetic evidence for a phonological rule: G-deletion in Turkish. *Research in Phonetics* 1: 217-232.
- Rueda-López, J. (2007). A plausible case of neutralization of /s/ and /r/ in Sevillian Spanish. *LL Journal* 2, 1: 1-13.
- Ruhm, H., E. Mencke, R. Milburn, W. Cooper & D. Rose (1966). Different sensitivity to duration of acoustic signals. *Journal of Speech and Hearing Research* 9: 371-384.
- Sabbagh, H. (2008). *Disambiguation and Relevance Theory in two Arabic dialects*. MA dissertation, Umm Al-Qura University.
- Sams, M., P. Paavilainen, K. Alho & R. Näätänen (1985). Auditory frequency discrimination and event-related potentials. *Electroencephalography and Clinical Neurophysiology* 62: 437-448.
- Samuel, A. (1982). Phonetic prototypes. *Perception and Psychophysics* 31, 4: 307-314.
- Sawyer, A. & J. Peter (1983). The Significance of statistical significance tests in marketing research. *Journal of Marketing Research* 20: 122-133.

- Scarr, S. (1997). Rules of evidence: A larger context for the statistical debate. *Psychological Science* 8: 16-17.
- Schafer, W. (1993). Interpreting statistical significance and nonsignificance. *Journal of Experimental Education* 61: 383-387.
- Schmidt, F. & J. Hunter (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. Harlow, S. Mulaik & J. Steiger (eds.), *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum Associates. 37-64.
- Schröger, E. & I. Winkler (1995). Presentation rate and magnitude of stimulus deviance effects on human pre-attentive change detection. *Neuroscience Letters* 193, 3: 185-188.
- Scobbie, J. & J. Stuart-Smith (2006). Quasi-phonemic contrast and the fuzzy inventory: Examples from Scottish English. *QMUC Speech Science Research Centre Working Papers WP-8*: 1-19.
- Scobbie, J., A. Turk & N. Hewlett (1999). Morphemes, phonetics, and lexical items: The case of the Scottish vowel length rule. *Proceedings of the XIVth International Congress of Phonetics Sciences* 2: 1617-1620.
- Scott, D. (1979). On optimal and data-based histograms. *Biometrika* 66: 605-610.
- Scott, D. (1992). *Multivariate density estimation: Theory, practice, and visualisation*. New York: John Wiley and Sons.
- Scott, D. (2009). Sturges' rule. *Wiley Interdisciplinary Reviews: Computational Statistics* 1, 3: 303-306.
- Sedlmeier, P., T. Betsch & F. Renkewitz (2002). Frequency processing and cognition: Introduction and overview. In P. Sedlmeier & T. Betsch (eds.), *ETC.: Frequency processing and cognition*. New York: Oxford University Press. 1-20.
- Shapiro, B. (1969). The subjective estimation of relative word frequency. *Journal of Verbal Learning and Verbal Behaviour* 8: 248-251.
- Shaw, R. & T. Mitchell-Olds (1993). Anova for unbalanced data: An overview. *Ecology* 74, 6: 1638-1645.
- Sheldon, D. (1973). A short experimental investigation of the phonological view of the writer-rider contrast in U.S. English. *Journal of Phonetics* 1: 339-346.
- Simonet, M., M. Rohena-Madrado & M. Paz (2008). Preliminary evidence for incomplete neutralization of coda liquids in Puerto Rican Spanish. In L. Colantoni & J. Steele (eds.), *Selected Proceedings of the 3rd Conference on Laboratory Approaches to Spanish Phonology*. Somerville, MA: Cascadilla Proceedings Project. 72-86.
- Singh, L. (2008). Influences of high and low variability on infant word recognition. *Cognition* 106, 2: 833-870.
- Skarnitzl, R. (2009). Challenges in segmenting the Czech lateral liquid. In A. Esposito & R. Vich (eds.) *Cross-Modal analysis of speech, gestures, gaze and facial expressions: Lecture notes in Computer Science 5641*: 162-172.
- Slowiaczek, L., & D. Dinnsen (1984). Neutralization and word-final devoicing in Polish. *Research on speech perception progress report No. 10*. Indiana University. 197-220.
- Slowiaczek, L., & D. Dinnsen (1985). On the neutralizing status of Polish word final devoicing. *Journal of Phonetics* 13: 325-341.

- Smith, J. (2001). Lexical category and phonological contrast. In R. Kirchner, J. Pater & W. Wikely (eds.), *PETL 6: Proceedings of the Workshop on the Lexicon in Phonetics and Phonology*. Edmonton: University of Alberta. 61-72.
- Smith, R. (2004). *The role of fine phonetic detail in word segmentation*. PhD thesis, University of Cambridge.
- Smolensky, P. (2006). Optimality in phonology II: Harmonic completeness, local contrast conjunction, and feature domain markedness. In P. Smolensky & G. Legendre (eds.), *The harmonic mind: From neural computation to Optimality-Theoretic grammar*. Cambridge, MA: MIT Press. 27-160.
- Smorodinsky, I. (2002). *Schwas with and without active gestural control*. PhD thesis, Yale University.
- Snodgrass, J. & J. Corwin (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology* 117: 34-50.
- Snodgrass, J., G. Levy-Berger & M. Haydon (1985). *Human experimental psychology*. New York: Oxford University Press.
- Snoeren, N., P. Hallé & J. Segui (2006). A voice for the voiceless: Production and perception of assimilated stops in French. *Journal of Phonetics* 34: 241-268.
- Solomons, L. (1900). A new explanation of Weber's Law. *Psychological Review* 7, 3: 234-240.
- Sommers, M. & J. Barcroft (2007). An integrated account of the effects of acoustic variability in 1st language and 2nd language: Evidence from amplitude, fundamental frequency, and speaking rate variability. *Applied Psycholinguistics* 28: 231-249.
- Sommers, M. & D. Kewley-Port (1996). Modeling formant frequency discrimination of female vowels. *Journal of the Acoustical Society of America* 99, 6: 3770-3781.
- Sommers, M., L. Nygaard & D. Pisoni (1994). Stimulus variability and spoken word recognition I. Effects of variability in speaking rate and overall amplitude. *Journal of the Acoustical Society of America* 96, 3: 1314-1324.
- Stager, C. & J. Werker (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature* 388: 381-382.
- Stein, R., E. Gossen & K. Jones (2005). Neuronal variability: noise or part of the signal? *Nature Reviews Neuroscience* 6: 389-397.
- Steriade, D. (1999). Phonetics in phonology: The case of laryngeal neutralization. In M. Gordon (ed.), *UCLA Working Papers in Linguistics* 2, *Papers in Phonology* 3: 25-246.
- Steriade, D. (2000). Paradigm uniformity and the phonetics-phonology boundary. In M. Broe & J. Pierrehumbert (eds.), *Papers in laboratory phonology V: Acquisition and the lexicon*. Cambridge: Cambridge University Press. 313-334.
- Sterne, J. & G. Davey Smith (2001). Shifting the evidence—what's wrong with significance tests? *Physical Therapy* 81, 8: 1464-1469.
- Stevens, K. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America* 111, 4: 1872-1891.
- Studebaker, G. (1985). A "Rationalised" Arcsine transform. *Journal of Speech and Hearing Research* 28: 455-462.

- Sturges, H. (1926). The choice of a class interval. *Journal of the American Statistical Association* 21, 153: 65-66.
- Tabain, M. & P. Perrier (2005). Articulation and acoustics of /i/ in pre-boundary position in French. *Journal of Phonetics* 33: 77-100.
- Tabain, M. & P. Perrier (2007). An articulatory and acoustic study of /u/ in preboundary position in French: The interaction of compensatory articulation, neutralization avoidance and featural enhancement. *Journal of Phonetics* 35, 2: 135-161.
- Taft, M. (2006). Orthographically influenced abstract phonological representation: Evidence from non-rhotic speakers. *Journal of Psycholinguistic Research* 35, 1: 67-78.
- Taft, M. & G. Hambly (1985). The influence of orthography on phonological representations in the lexicon. *Journal of Memory and Language* 24: 320-335.
- Tanner, W. (1961). Physiological implications of psychophysiological data. *Annals of the New York Academy of Sciences* 89: 752-765.
- Tatham, M. (1976). Variability in phonetics. *York Papers in Linguistics* 6: 47-53.
- Taylor, D. (1975). The inadequacy of the bipolarity and distinctive features: The German 'voice-voiceless' distinction. In P. Reich (ed.), *The 2nd LACUS Forum*. Chicago: Chicago Linguistic Circle. 107-119.
- Thalheimer, W. & S. Cook (2002). *How to calculate effect sizes from published research articles: A simplified methodology*. Retrieved 20 October 2008 from http://work-learning.com/effect_sizes.html.
- Thompson, B. (1989). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. *Measurement and Evaluation in Counseling and Development* 22: 2-6.
- Thompson, B. (1998a). Statistical significance testing and effect size reporting: Portrait of a possible future. *Research in the Schools* 5, 2: 33-38.
- Thompson, B. (1998b). In praise of brilliance: Where that praise really belongs. *American Psychologist* 53: 799-800.
- Thoresen, C. & J. Elashoff (1974). "An analysis-of-variance model for intrasubject replication design": Some additional comments. *Journal of Applied Behavior Analysis* 7, 4: 639-641.
- Thurstone, L. (1927). A mental unit of measurement. *Psychological Review* 34: 415-423.
- Tieszen, B. (1997). *Final stop devoicing in Polish: An acoustic and historical account for incomplete neutralization*. PhD thesis, University of Wisconsin, Madison.
- Tiitinen, H., P. May, K. Reinikainen & R. Näätänen (1994). Attentive novelty detection in humans is governed by pre-attentive sensory memory. *Nature* 372: 90-92.
- Tilkov, D. (1982). *Gramatika na Sâvremennija Bâlgarski Ezik: Fonetika*. Sofia: Izdatelstvo na BAN.
- Todd J. & P. Michie (2000). Do perceived loudness cues contribute to duration mismatch negativity (MMN)? *NeuroReport* 11, 17: 3771-3774.
- Topintzi, N. (2006). *Moraic onsets*. PhD thesis, University College London.

- Torres, N. (2001). *Voicing assimilation in Catalan and English*. PhD thesis, Universitat Autònoma de Barcelona.
- Traunmüller, H. (1988). Paralinguistic variation and invariance in the characteristic frequencies of vowels. *Phonetics* 45: 1-29.
- Traunmüller, H. (1994). Conventional, biological and environmental factors in speech communication: A modulation theory. *Phonetica* 51: 170-183.
- Trubetzkoy, N. (1969). *Principles of phonology*. Translated by C. Baltaxe. London: University of California Press.
- Trudgill, P. (1974). *The social differentiation of English in Norwich*. Cambridge: Cambridge University Press.
- Trumpower, D. & O. Fellus (2008). Naïve statistics: Intuitive analysis of variance. *Proceedings of the 30th Annual Conference of Cognitive Science Society*. 499-503.
- Tryon, W. (1998). The instructable null hypothesis. *American Psychologist* 53, 7: 796.
- Tryon, W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods* 6, 4: 371-386.
- Tsukada, K. (2009). An acoustic comparison of vowel length contrasts in Arabic, Japanese and Thai: Durational and spectral data. *International Journal of Asian Language Processing* 19, 4: 127-138.
- Tufte, E. (1997). *Visual explanations: Images and quantities, evidence and narrative*. Cheshire, Connecticut: Graphics Press.
- Tukey, J. (1977). *Exploring data analysis*. Reading, Mass: Addison-Wesley.
- Tukey, J. (1991). The philosophy of multiple comparisons. *Statistical Science* 6: 100-116.
- Tuller, B., P. Case, M. Ding & J. Kelso (1994). The nonlinear dynamics of speech categorization. *Journal of Experimental Psychology: Human Perception and Performance* 20, 1: 3-16.
- Tuller, B., M. Jantzena & V. Jirsa (2008). A dynamical approach to speech categorization: Two routes to learning. *New Ideas in Psychology* 26: 208-226.
- Turk, A., S. Nakai & M. Sugahara (2006). Acoustic segment durations in prosodic research: A practical guide. In S. Sudhoff, D. Lenertová, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter & J. Schließer (eds.), *Methods in empirical prosody research*. Berlin: Walter de Gruyter. 1-27.
- Turk, A. & J. Sawusch (1996). The processing of duration and intensity cues to prominence. *Journal of the Acoustical Society of America* 99, 6: 3782-3790.
- Tversky, B. (2005). Prolegomenon to scientific visualizations. In J. Gilbert (ed.), *Visualization in science education*. 29-42.
- Tversky, B., J. Zacks & B. Martin (2008). The structure of experience. In T. Shipley & J. Zacks (eds.), *Understanding events: From perception to action*. Oxford: Oxford University Press. 436-464.
- Utman, J., S. Blumstein & M. Burton (2000). Effects of subphonemic and syllable structure variation on word recognition. *Perception and Psychophysics* 62, 6: 1297-1311.

- Välismaa-Blum, R. (2009). The phoneme in cognitive phonology: Episodic memories of both meaningful and meaningless units? *CogniTextes 2*. Retrieved on 24th March 2010 from <http://cognitextes.revues.org/211>.
- Vaux, B. & B. Samuels (2005). Laryngeal markedness and aspiration. *Phonology 22*: 395–436.
- Vicente, K. & G. Torenvliet (2000). The earth is spherical ($p < 0.05$): Alternative methods of statistical inference. *Theoretical Issues in Ergonomics Science 1, 3*: 248–271.
- Wang, W. & J. Fillmore (1961). Intrinsic cues and consonant perception. *Journal of Speech and Hearing Research 4*: 130–136.
- Warner, N., E. Good, A. Jongman & J. Sereno (2006). Orthographic vs. morphological incomplete neutralization effects: Letter to the editor. *Journal of Phonetics 34*: 285–293.
- Warner, N., A. Jongman, J. Sereno & R. Kemps (2004). Incomplete neutralization and other sub-phonemic durational differences in production and perception: Evidence from Dutch. *Journal of Phonetics 32*: 251–276.
- Wayland, R. & S. Guion (2003). Perceptual discrimination of Thai tones by naïve and experienced learners of Thai. *Applied Psycholinguistics 24*: 113–129.
- Wayland, S., J. Miller & L. Volaitis (1994). The influence of sentential speaking rate on the internal structure of phonetic categories. *Journal of the Acoustical Society of America 95, 5*: 2694–2701.
- Weisberg, H. (1992). *Central tendency and variability*. Newbury Park, CA: Sage Publications.
- West, P. (1999). The extent of coarticulation of English liquids: An acoustic and articulatory study. *Proceedings of the 14th International Congress of Phonetic Sciences*. San Francisco. 1901–1904.
- West, P. (2000). Perception of distributed coarticulatory properties of English /l/ and /a/. *Journal of Phonetics 27*: 405–425.
- Westbury, J. & P. Keating (1986). On the naturalness of stop voicing, *Journal of Linguistics 22*: 145–166.
- Whalen, D. (1991). Infrequent words are longer in duration than frequent words. *Journal of the Acoustical Society of America 90, 4*: 2311.
- Whalen, D. (1992). Further results on the duration of infrequent and frequent words. *Journal of the Acoustical Society of America 91, 4*: 2339–2340.
- Whalen, D., C. Best & J. Irwin (1997). Lexical effects in the perception and production of American English /p/ allophones. *Journal of Phonetics 25, 4*: 501–528.
- Whalen, D. & A. Levitt (1995). The universality of intrinsic F0 of vowels. *Journal of Phonetics 23*: 349–366.
- van Wijngaarden, S., H. Steeneken & T. Houtgast (2002). Quantifying the intelligibility of speech for non-native listeners. *Journal of the Acoustical Society of America 111, 4*: 1906–1916.
- Winitz, H., M. Scheib & J. Reeds (1972). Identification of stops and vowels from the burst portion of /p, t, k/ isolated from conversational speech. *Journal of the Acoustical Society of America 51, 4*: 1309–1317.

- Wissing, D. & A. van Rooy (1992). Onvolledige neutralisatie: die ontstemmingsreël in Afrikaans. *South African Journal of Linguistics, Suppl. 13*: 135-144.
- Wittels, P., B. Johannes, R. Enne, K. Kirsch & H-C. Gunga (2002). Voice monitoring to measure emotional load during short-term stress. *European Journal of Applied Physiology 87*: 278-282.
- Wood, C. (1974). Parallel processing of auditory and phonetic information in speech perception. *Perception and Psychophysics 15*: 501-508.
- Wood, S. (1996). Assimilation or coarticulation? Evidence from the temporal coordination of tongue gestures for the palatalization of Bulgarian alveolar stops. *Journal of Phonetics 24*: 139-164.
- Wood, S. & T. Pettersson (1988). Vowel reduction in Bulgarian: The phonetic data and model experiments. *Folia Linguistica 22*: 239-262.
- Wouters, J. & M. Macon (2002). Effects of prosodic factors on spectral dynamics II: Synthesis. *Journal of the Acoustical Society of America 111, 1*: 428-438.
- Xu, F. (2003). Numerosity discrimination in infants: Evidence for two systems of representations. *Cognition 89, 1*: B15-B25.
- Xu, F. & E. Spelke (2000). Large number discrimination in 6-month-old infants. *Cognition 74*: B1-B11.
- Xu, F., E. Spelke & S. Goddard (2005). Number sense in human infants. *Developmental Science 8, 1*: 88-101.
- Yip, M. (1996). Lexicon optimization in languages without alternations. In J. Durand & B. Laks (eds.), *Current trends in phonology: Models and methods II*. Salford: ESRI, University of Salford Publications. 757-788.
- Yip, M. (2002). *Tone*. Cambridge: Cambridge University Press.
- Ylinen, S. (2006). *Cortical representations for phonological quantity*. PhD thesis, University of Helsinki.
- Yoo, I. & B. Blankenship (2003). Duration of epenthetic /t/ in polysyllabic American English words. *Journal of the International Phonetic Association 33*: 153-164.
- Yu, A. (2007). Understanding near mergers: The case of morphological tone in Cantonese. *Phonology 24*: 178-214.
- Yu, A. (2009). Contrast and neutralisation. Paper presented at *the EALing Fall School in Linguistics*, Département des études cognitives, École Normale Supérieure, 14-22 September 2009, Paris.
- Yun, G. (2007). Word frequency, stress and coarticulation in English. *Studies in Phonetics, Phonology and Morphology 13, 2*: 315-331.
- Yun, W. (2008). Noun-verb asymmetries in Korean phonology. In N. Abner & J. Bishop (eds.), *Proceedings of the 27th West Coast Conference on Formal Linguistics*. Somerville, MA: Cascadilla. 449-457.
- Zacks, R. & L. Hasher (2002). Frequency processing: A twenty-five year perspective. In P. Sedlmeier & T. Bestch (eds.), *ETC.: Frequency processing and cognition*. New York: Oxford University Press. 21-36.
- Zacks, J. & B. Tversky (2001). Event structure in perception and conception. *Psychological Bulletin 127*: 3-21.
- Zacks, J., B. Tversky & G. Iyer (2001). Perceiving, remembering, and communicating structure in events. *Journal of Experimental Psychology: General 130*: 29-58.

- Zawaydeh, B. (1999). *The phonetics and phonology of gutturals in Arabic*. PhD thesis, Indiana University.
- Zhang, J. (2007). Constraint weighting and constraint domination: A formal comparison. *Phonology* 24: 433-459.
- Zue, V. & M. Laferriere (1979). Acoustic study of medial t, d in American English. *Journal of the Acoustical Society of America* 66: 1039-1050.
- Zurairq, W. & J. Sereno (2007). English lexical stress cues in native English and non-native Arabic speakers. *International Congress of Phonetic Sciences XVI*. Saarbrücken 6-10 August 2007. 829-832.